Ethical DNA Model for Artificial General Intelligence

Ralph C. Ennis LEAD Consulting, 9418 Harvest Acres Court, Raleigh, NC 27617 ralphennis@gmail.com

Abstract. An effort to understand ethical reasoning must not focus on a list of ethical rules but the underlying grammar, an ethical DNA, for the development of all ethical precepts. The purpose of this paper is to put forth a framework for ethical DNA (e-DNA) in a manner that is applicable to the pursuit of artificial general intelligence (AGI). This e-DNA model revolves around nine continuums and their intersections and interactions. The generality of any ethical DNA model is suggested only as it shows utility across cultural diversity. With the use of this e-DNA, the Japanese construct of amae is parsed. Amae is a complex construct within the Japanese society that impacts human relations—and thus ethical behavior among relations. The utility of the e-DNA model for artificial intelligence is evident in the geometric interactions between the continuums that provide a way forward in programming.

Keywords. Ethics, moral, artificial general intelligence, decision model

1 Introduction

Mikhail (2007) frames the following poignant question relevant in the pursuit of an ethically-based artificial general intelligence (AGI): "Is there a universal moral grammar and, if so, what are its properties?" Stated otherwise, is there a set of rules that govern the formation of all ethically acceptable behaviors across cultures?

Evidence can be found on any kindergarten playground across the global community that ethical reasoning is at play. In what part of the human experience is some construct of "fairness and harmony" non-existent? This construct may seem suspended or violated at various times, but an innate awareness of fairness and harmony resides within us all—even in our early childhood interactions (Smith, et al., 2013).

Fairness may be defined differently across individuals, families and cultures, but yet it resonates within all social structures even if pathways to it are blocked. Fairness to some implies non-bias equality of quantity and quality. However, this definition rarely works out well without the consideration of context.

For instance, is it fair to an eight year-old sister to be treated equally with her four year-old brother, or vice-a-versa? Most parents would conclude unequal treatment is far more "fair" that an unwavering pursuit of equality. Much to the consternation of young siblings, most parents conclude that it does not have to be equal to be fair. Fairness is contextual to age, abilities, available resources, etc.

If fairness is not somehow achieved or at least approximated, we humans recognize that harmony (dynamic balance) within a system may be threatened or disrupted.

Back to the family system—sibling disputes over fairness can disrupt the sense of harmony for all in the family.

What remains in the pursuit of ethical reasoning is not the question of a set of ethical rules that are proven to be universal, but rather can a grammar—a functional ethical DNA be established? By using that DNA of ethical reasoning, can a diversity of contextual rules be fashioned and situations evaluated for ethical acceptability? Is that DNA applicable in the formation of ethical rules and parsing of existing rules across cultures—even when the rules seem in conflict?

A solution to that ethical DNA (e-DNA) and subsequent management of it is paramount in the quest for artificial general intelligence (AGI) (Gubrud, 1997). This e-DNA should account for the human sense of fairness and harmony across a multitude of contexts. Asimov (1950) proposed such a moral code with his three laws of robotics, but we need a more fundamental code from which these laws and others might be derived. As Pana (2006) states, "We do not have to implement a moral code, but to create a moral intelligence, we can aspire to a condition of potentiality, not the generation of some fixed reality."

In this paper, I will posit an e-DNA model that has applicability across cultures and is adaptable to AGI. This e-DNA will account for human ethical reasoning and allow for such reasoning at a machine level of intelligence.

In short, the e-DNA code involves nine continuums subdivided as logic of intellect, logic of emotion and imagined outcomes. These nine continuums are considered in this paper along with three central constructs that arise from their intersections and interactions. These continuums allow for gradation of each endpoint on a linear scale. Furthermore, the logic of intellect, logic of emotions and imagined outcomes axes are non-hierarchal. All are conceptualized with equal weight in the decision making process.

2 Continuums and Central Constructs for e-DNA

The twentieth century European philosopher Edwin Wittgenstein stated: "Language is a labyrinth of paths. You approach from one side and know your way about; you approach the same place from another side and no longer know your way about "(Philosophical Investigations 203). With this labyrinth in mind, the e-DNA model is established "on continuums" rather than separate factors.

Though this approach is debatable, much of ethical reasoning fails to fit neatly within discrete categories. Humanity devises detailed laws to fulfill that sense of discrete ethical boundaries. However, even then the need for the "spirit of the law" to triumph the "letter of the law" becomes situationally mandatory in order to prioritize laws. For instance, the letter of highway laws may state a certain speed limit. However, if one needs to go a little faster to secure the life of a person with a medical emergency and without jeopardizing the life of other drivers, then most would conclude that some bending of the letter of the law (speed limit) to preserve the spirit of the law (preservation of life) is ethically sound reasoning.

The language of e-DNA will be put forth in English. However, each of the nine continuums can be translated into most languages with some degree of accuracy. This language difference must be accounted for—but not at this point. The nine continu-

ums are grouped in three broader categories (see Figure 1): logic of intellect, logic of emotion and imagined outcomes. Each line in three-dimensional Euclidean space represents a continuum. Logic of intellect refers to the common language notion of "thinking a matter through to a conclusion without emotional bias". Logic of emotions comes into play when feelings, molded by cultural interpretations into emotional constructs, impact the logic of intellect. And finally, imaginations of probabilistic outcomes impact and adjust our intellect and emotional logics. The arrows in Figure 1 point to the intersection of three continuum which form a central construct for the logics and imagined outcomes.

For example, a society may disqualify a Judge from trying a suspected murderer of the son of that Judge. There is a high probability that the emotions of the Judge will blind him from conducting due process of law driven by a logic of intellect. Furthermore, the imagined outcome of such a trial will not serve the cause of justice among members in a society.

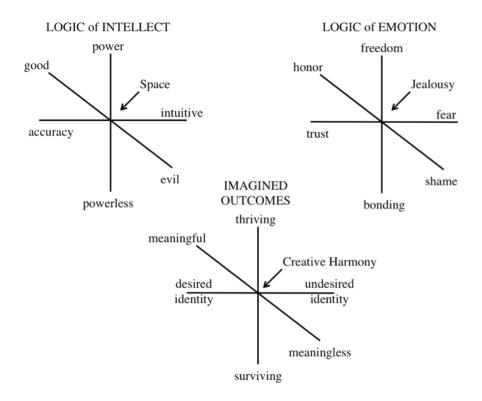


Figure 1: Ethical DNA Continuums

2.1 Ethical Logic of Intellect

Good-Evil. The very fact that all cultures have some sense of good and evil, even though they may disagree on the details, is an indication of an e-DNA. Thus a good-

evil continuum simply states the obvious—humans think about ethics. This continuum must be included even if it seems redundant. However, it isn't enough to say that humans logics about good and evil. More goes into ethical reasoning than a final assessment of good and evil.

Accuracy-Intuitive. Morality includes of verbal and non-verbal truth telling that is accurate to facts or intuitively consistency with the facts. Many courts of law require witnesses to vow to tell the truth. This sense of accuracy from one's point of view is fundamental to ethical reasoning even as multiple points of view better shape an accurate account of a situation.

Powerful-Powerless. Care of the powerless, e.g. young children, is foundational to the continuation of any society. Moreover, the Hippocratic oath in medicine reasons along these lines of power. Its core tenant is to do good to patients and never harm them. This "good" is reasoned as prescribing procedures and substances to bring about better health. Better health is wrapped in the concept of power and doing harm implicitly decreases this power on a continuum to death (i.e. total powerlessness).

Space. Mental or physical spatial ownership (individual and/or corporate) is the central construct of logic of intellect. Thus space can be conceptualized as "good," "evil," "accurate," intuitive," "powerful," and "powerless." Many wars (an ethically entangled pursuit) have and continue to be fought over some conflict of space.

2.2 Ethical Logic of Emotion

Freedom-Bonding. As ethical reasoning, the continuum of freedom-bonding is best understood at the extremes of abandonment and bondage. For a parent to be totally free without any attachments is viewed by society as abandoning their child to others or to society. To be in bondage suggest varying degrees and forms of slavery. However, healthy bonding and various levels of freedoms are necessary for individuals and societies.

Honor-Shame. The management of moral behavior often comes through positive rewards that honor people or negative consequences that shame them. Sometimes the concept of authority is embedded with honor and shame. Shaming is a common form of reforming deviate behavior at home, in the classroom as well as in the broader society. Thus shame remains as an endpoint of this continuum that is the hoped for (by authorities) emotional consequence of unethical behavior. The feeling of guilt is often linked to shame. Guilt indicates lapses in behavior; shame indicates remorse in one's identity (Lewis, 1995).

Trust-Fear. A breech of trust is often considered an ethical failure. Legal contracts are formed to fortify and ensure verbal trust. Fear of the consequences of broken trust often helps negotiate trust relationship.

Jealousy. Jealousy is posited as the central construct of the logic of emotion. Jealousy has two sides – jealous for and jealous of. The latter is better referred to as envy (Clanton, 1998). To cease to be jealous for someone that relies on that jealousy for their protection can constitute a breach of ethics. For example, marriage is a relationship fraught with jealousy—preferable jealousy for, not jealousy of. At its best, jealousy for involves an emotional bonding that brings freedom, a sense of honor between members and a trust that exist when members are present or apart. At its worst,

jealousy of can divide and destroy relationships. Furthermore, jealousy is seen to be ubiquitous in human cultures by Johnson and Price-Williams (1996).

2.3 Imagined Outcomes

Desired-Undesired Identity. To violate a person's identity through some abuse often causes strong negative reactions. Human identity structures are many and far reaching on their impact of ethical behavior. Wars have been fought to protect or advance national identities. Family inheritance laws fortify family identities within society. Certain professional identities improve the probability of securing research grants. And the imagined outcomes of present actions impact one's desired identity while decreasing the chance of an undesired identity.

Thriving-Surviving. The ethics of thriving hopefully does not value the elimination of others' survival. Humanity seeks to survive and from that basis thrive. The construct of thriving is highly imaginative. For instance, thriving in one cultural context may be imagined as possessing a cow or a bicycle. In another culture, those possessions might represent a subsistence survival.

Meaningful-Meaningless. Philosophy, art, religions are manifestation of humanity's quest for meanings that transcend themselves. Humanity, for the most part, imagines itself to be meaningful. Meaningless is conceptualized as a disruptor of productive living (thus interfering with the pursuit of thriving). Belief and aesthetic systems are designed to bring meaning into the human experience from conception to death. To violate these meanings can be considered an immoral act. Wars have been and continue to be fought over meanings, especially religious and political meanings.

Creative Harmony. The central construct of imagined outcomes is creative harmony. This ethical concept helps maintain the goodness of perpetrating harmonious health in individuals, enterprises and societies. The violation of creative harmony—destructive dissonance—can be viewed as morally wrong under certain but not all circumstances. Civil disobedience usually seeks a better long-term creative harmony in society through a short-term pathway of destructive dissonance to reshape the rules of society. Further explanations of these continuums are put forth by Ennis (2004).

3 Central Constructs of the Continuums

The uncommon word set "creative harmony of jealous space" is achieved by overlapping the central constructs of logic of intellect, logic of emotion and imagined outcomes (see Figure 2). Ethical reasoning implies each of these ideas. Jealous space allows for property rights; both physical and mental space is inherent in the spatial-temporal nature of language. Without jealousy a sense of possession and ownership, that pervades ethical reasoning, would be a mute issue. Thus we return to the ideas of "fairness" and "harmony" in systems. The negotiation of jealousy across spatial constructs will account for "fairness" and "fairness" is mediated through "harmony" that is dynamic and thus creating new states of being across time and space.

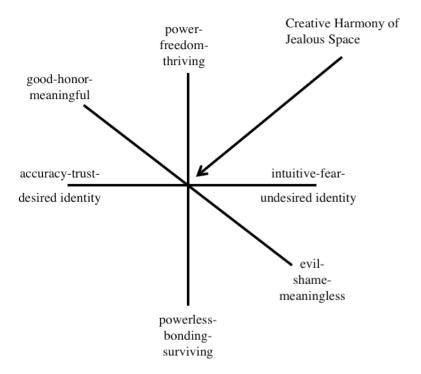


Figure 2: Overlapping Ethical DNA Continuums

4 Evaluating the Utility of e-DNA through Beauty and Love

In an effort to move from uncommon language to ordinary language, a discussion of ethical reasoning from the lens of "beauty" and "love" is needed. The previously discussed words "harmony" and "fairness" (which are foundational to ethical good) can be conceptualized as pathways to "beauty" and from "love." Toward a pursuit of harmony, a system acquires a sense of beauty. And from a motivation and commitment of love, acts of fairness, that do not necessarily achieve equality, are ethically optimal.

Beauty has been much debated through the millennia. The ancient Greeks consider it one of the three hallmarks of philosophy along with truth and goodness. "Truth" has been embedded within the e-DNA model as accuracy that can be trusted to present an identity that is mutually agreed upon ("desired"). "Goodness" is seen as the DNA of ethical reasoning that included both the good-evil reasoning continuum as well as the full nine overlapping continuums interactively engaged. But "beauty" must be unpacked more intuitively.

The culturally impacted construct of beauty yields a broad diversity that must account for tastes in spatial presentations (e.g. clothing fashions, facial shapes), charac-

ter generalizations (e.g. virtuous character and beautiful personality) and even beauty in power disruptions (e.g. distant stars forming and exploding). Without a sense of beauty and its opposite, ugliness, ethical reasoning might degenerate to only quantitative measurements of "fairness" and "balance." However, humanity's attraction to beauty and aversion to the violation of beauty (resulting in ugliness) makes ethical reasoning far more interesting and problematic.

Beauty can be perceived through the lens of creative harmony of jealousy space. Beauty can be conceived as displaying creative harmony amidst space that is jealousy held. Similarly, ugliness can be posited as displaying disharmonious jealous space. The good and evil of beauty and ugliness is thus a matter of negotiating jealous space.

The link between beauty and jealous space is intuitive. Beauty reveals an emotional attraction focused on some spatial object or spatially grounded concepts (such as symmetry). An attraction can be conceptualized as a jealousy—a desire to possess for oneself (at some distance) that which is deemed precious. Space that is jealously possessed and is in creative harmony with other jealously held spaces may be deemed beautiful within a family, a business system, a culture. However, when a space is jealousy possessed by conflicting parties, these jealousies (i.e. destructive envies) can produce an ugliness that can lead to brutal conflicts. Thus the underlining dynamics of jealous space is intrinsically embedded within human reasoning of beauty-ugliness.

This beauty is on a continuum with ugliness. Degrees of beauty are compared with degrees of ugliness. Consistent with the above definition of beauty, ugliness is posited as the violation of creatively harmonious jealous space—thus disharmony of envied space. The comparative difference is primarily within the definitions of jealousy and envy. Jealous is a jealousy "for" something or someone with an established right of ownership, while envy is a jealousy "of" something or someone with no established right of ownership. (Obviously, establishing rightful ownership can be problematic.)

For instance, societies agree that parents have some limited right of ownership to their children. For a parent to be jealous "for" the space of his/her child is a beautiful act of harmony. However, when a parent becomes jealous "of" (envious of) the child, something very different occurs, something very ugly. To be jealous "of" is an intrusion of personal space. Parental jealousy "for" can nurture the child while envy, jealousy "of", can rob the child of the space necessary for protection and development.

The desire (and sometimes an act) to invade the space of another and take from him/her that which he rightfully possesses is an ugliness that humanity is acquainted with. This envy, this over-possessive, misdirected and deformed jealousy, can undermine human relationships while a proper sense of jealousy "for" another can help protect and develop a person who is cherished within that possessive jealousy.

For example, if one is jealously possessives of his/her own sexual space (body) and someone attempts to enter that body space without permission, then an internal emotional reaction will occur indicating that this intrusion is an unfair violation, that this act is an ugliness warranting the labels of "evil". Thus, it is culture-general to discuss and condemn the destructive ugliness of sexual rape. Rapes in wartime have sometimes been justified throughout history as acts of conquering the enemy. Fortunately, such wartime violations of jealous space are condemned by the Geneva Convention

Another common word associated with ethical reasoning is "love". Love is determined by individuals and societies to be both a high ideal and a base passion. Love as

an ethic is nebulous. Love may motivate many ethical pursuits. Moreover, the absence of love, when love is expected or longed for, or the presence of hate (love's opposite), invokes ethical choices. Love can be conceptualized as an internal working of beauty and for beauty. And beauty, creative harmony of jealous space, is an outward evidence of some love in action. Furthermore, love as a motive helps mitigate the necessity of fairness that is not always equal.

A final example of the utility of "creative harmony of jealous space" that defines beauty is a tragedy of ugliness and evil. Cruel ugliness reigned in the Rwandan genocide of 1994 in which an estimated 800,000 people were killed in 100 days. One people group, the majority Hutus, sought creative harmony for their desired identity by denigrating their opposition as "cockroaches" (an undesired identity) and systematically labored to eradicate them. They negotiated their space (i.e. their country with physical land and property) with a jealousy that became envious, over-possessive and oppressive to the minority Tutsi population. This negotiation of jealous space allowed a justification for the evils of genocide—a justification acceptable at that time to many (not all) Hutus while being totally unacceptable to all Tutsi. Thus, the e-DNA model can be used in parsing highly charged and ethically implicit behaviors that are disastrously ugly and evil.

The construct of beauty as creative harmony of jealous space holds promise as an e-DNA in negotiating the abstract and practical ethical discussions of our day across cultural distinctions. In going forward, an analysis of ethical reasoning patterns across cultures is needed. This analysis can serve to reinforce the case for this e-DNA model driven by beauty as creative harmony of jealous space.

5 Generality of e-DNA Suggested

The e-DNA model is a generalization that can be useful across various a wide variety of cultural setting. From this generalized model, differences from culture, gender, age, etc. that are prevalent in ethical reasoning can be derived. Generality is suggested through five perspectives.

First, the concept of "creative" can be viewed as a generalization since "change over time" (necessary for creativity) is inherent in all ethical systems of thought—even as language itself changes over time. Second, "harmony" can be perceived as a general ethical construct since its complete opposite insures annihilation of any set of identities (e.g. the destruction of civilizations). Third, jealousy can be projected across cultures from the play of jealousy within the Oedipus complex that has been documented in over 100 cultures (Johnson, A. W. & Price-Williams, D., 1996). Fourth, spatial constructs are inherent in all language at various level of abstraction. Language development starts with objects (e.g. "mommy"), usually associate with some time marker and then over time generalizations and abstractions are formed that make transmission of meanings between persons a fruitful enterprise.

The fifth perspective for suggesting generality will be a specific parsing of a Japanese word, *amae*, using the e-DNA model that has been put forth in English (see Table 1). This cross-cultural evaluation will contribute evidence for the generality of the model.

| e-DNA Model | Japanese Amae Parsed |
|---------------------------------------|---|
| Logic of Intellect | supunese ilmue i urseu |
| Powerful – powerless | Amae requires the powerlessness of receiving as a child would and yields the power of being provided for. |
| Good – evil | Amae requires an acknowledgement of good in one's in-group and holds that evil is betrayal of one's in-group. |
| Accuracy – intuition | Amae requires intuition to negotiate relationships and assumes the accurate interpretation of amae as a social construct. |
| Space | Amae requires the negotiation of space between two or more people. |
| Logic of Emotion | |
| Trust – fear | Amae requires trust in other(s) and it implies the fear of being betrayed by others. |
| Honor – shame | Amae requires the honor of submitting to another's will and it forbids the shame of betraying another. |
| Freedom – bonding | Amae requires the bonding of dependency and yields the freedom of dependency. |
| Jealousy | Amae requires the management of a privileged and thereby jealous relationship between people. |
| Imagined Outcomes | |
| Surviving – thriving | Amae views the proper networking of relationships for both surviving and thriving. |
| Desired identity – undesired identity | Amae views self as dependent as a desired identity and views the absence of a dependent relationship as an undesired identity. |
| Meaningful – meaningless | Amae views the parent-child relationship as the fundamental meaningful relationship and the absence of amae as fundamentally a meaningless existence. |
| Creative harmony | Amae requires both persons in an amae relationship maintain and creatively enhance harmony |

Table 1: Using e-DNA Model to Parse the Japanese Amae Construct

Japanese psychiatrist Takeo Doi (1981) described in detail the dynamics of *amae* in the Japanese culture stating, "The Japanese term *amae* refers, initially, to the feelings that all normal infants at the breast harbor toward the mother – dependence, the desire to be passively loved, the unwillingness to be separated from the warm mother-child circle and cast into a world of objective 'reality' " (p. 7). He went on to say, "... all the many Japanese words dealing with human relations reflect some aspect of the *amae* mentality. This does not mean, of course, that the average man is clearly aware of *amae* as the central emotion in ninjo (human feeling)" (p. 33).

Regarding the impact of *amae* on the culture, he stated, "Only a mentality rooted in *amae* could produce a people at once so unrealistic yet so clear-sighted as to the basic human condition; so compassionate and so self-centered; so spiritual and so materialistic; so forbearing and so willful; so docile and so violent" (p. 9).

Furthermore, Doi compared the Japanese with Westerners in stating, "Scholars have put forward many different theories concerning the ways of thinking of the Japanese, but most agree in the long run that, compared with thought in the West, it is not logical but intuitive" (p.76). Doi proposed outsiders struggle with the *amae* construct. He stated, "... to persons on the outside who do not appreciate *amae* the conformity imposed by the world of *amae* is intolerable, so that it seems exclusivist and private, or even egocentric" (p. 77).

The e-DNA model analysis of the Japanese construct of *amae* is not intended to fulfill the richness of the Japanese construct but rather to approximate its construction in such a way that the multi-variable applications of *amae* may be anticipated and appreciated within the Japanese cultural context. This analysis of *amae* contributes evidence for the generality of the e-DNA model across human cultures.

6 Using e-DNA in AGI

In hierarchical structures, one would need to prioritize the three proposed central constructs of e-DNA. However as previously mentioned, Wittgenstein suggested "Language is a labyrinth of paths" (Philosophical Investigations 203). This e-DNA model, with overlapping and interacting continuums, accounts for the inherent convolutions—labyrinth of paths—of common human language without establishing a true hierarchy among the central constructs.

Earlier the question arose of accounting for differences in language translations of the words used on the continuums. The labyrinth of paths in language helps alleviate this problem. The assumption that language is discrete and static requires fixed constants that provide exact translations rather than variables within an approximated range. (This range does not allow non-sensical relativism that would cancel the prospect of transference of meanings). This e-DNA model opts for an approximated range of meanings.

The geometrical structures of the e-DNA model lend themselves to computer programming. This set of (3) 3-D grids provides an acceptable means for mapping ethically constructs.

By parsing (with the inputter's bias accounted for) an abundance of words in sentence and image contexts, a more general understanding of the ethical use of a word

can be extrapolated. This extrapolation can then be used in evaluating and/or forming ethical rules of thumb. That ethical evaluation would be on a continuum from optimal, acceptable, warning to dangerous.

This e-DNA model can evaluate and suggest optimizing pathways for the richness of ethical reasoning required for true AGI. And without which the imagined outcome of super-human artificial intelligence can only be seen as devastating for humanity. If AGI machines advance with only an ethic of effectiveness and efficiency (inherent in almost all programming), then thriving and surviving might well dominate the struggle between humanity and machine in the decades ahead—with machine the predictable winner.

7 Conclusion

This paper put forth a means of describing an ethical DNA and illustrated its utility in parsing an ethically implicit Japanese construct. In seeking to establish an e-DNA model, I have posited nine overlapping and interacting continuums with three central constructs. Evidence for its generality has been provided.

If human-level artificial intelligence is to be achieved, the DNA code of thought and behavioral decisions must also be articulated and translated into machine language process and output decisions. Decisions are foundational to human intelligence. The human mind seems to parse all decisions in a seamless fashion while seeking congruence and abating dissonance. This parsing process is mostly opaque to us all. Describing process (thought) decisions and output (behavioral) decisions are essential for achieving human-level artificial intelligence.

This paper stops short of the point where usage of e-DNA can account for all process and output decisions that human-level intelligence achieves. Future work can extend the e-DNA model to account for this full range of decisions needed for AGI.

References

Asimov, I.: I, Robot, New York: Doubleday & Company. (1950)

Clanton, G.: A sociology of jealousy. In G. Clanton & L. G. Smith (Eds.). *Jealousy* (3rd ed.). New York, NY: University Press of America. (1998) 297-312

Doi, T.: *The anatomy of dependency: The key analysis of Japanese behavior*. Tokyo, Japan: Kodansha International (1981).

Ennis, R.: A theoretical model for research in intercultural decision making. *Intercultural Communication Studies*. 8: (2004) 113-124

Gubrud, M.: Nanotechnology and International Security, *Fifth Foresight Conference on Molecular Nanotechnology* (November 1997)

Johnson, A. W. & Price-Williams, D.: *Oedipus ubiquitous: The family complex in world folk literature.* Stanford, CA: Stanford University Press (1996)

Lewis, H. Shame, repression, field dependence, and psychopathology. In *Repression and dissociation: Implications for personality theory, psychopathology and health.* Chicago, IL: University of Chicago Press. (1995) 239-241

Mikhail, J.: Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*. 11(4) 143-152 (2007)

Pana, L.: Artificial intelligence and moral intelligence. TripleC 4(2) 254-264 (2006)

Smith, C., Blake. P., Harris, P.: I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *Plos One* 10.1371 (2013)

Wittgenstein, L.: *Philosophical Investigations*. Trans. G. E. M. Anscombe. Oxford: Blackwell Publishers. (1999)

Referencing this paper:

Ennis, R. An Ethical DNA Model for Artificial General Intelligence. USB Proceedings The 10th International Conference on Modeling Decisions for Artificial Intelligence. MDAI 2013, Barcelona, Catalonia November 20 - 22, 2013, Pages 56 -67. ISBN: 978-84-695-9120-8 D.L.: B. 27203-2013