A Model for Artificial Superintelligent Agents to Reason Ethically Across the Religious Persuasions of Humanity

Ralph C. Ennis

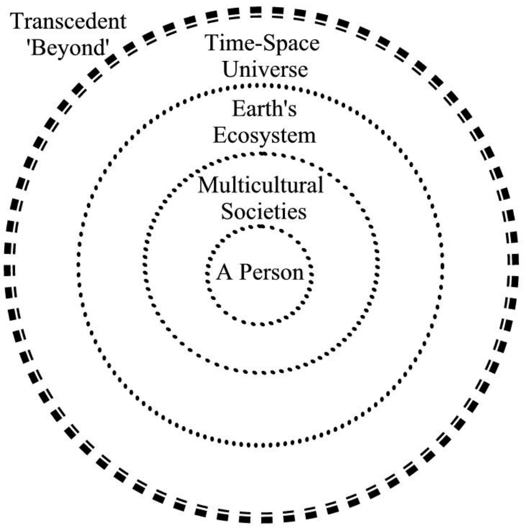
The default ethical reasoning of an artificial superintelligent agent (SIA) will be the effective and efficient acquisition of power in its many forms. In short, "get it done—fast." This goal seek approach is highly advantageous within delimited problems such as eliminating all forms of cancers, feeding 15 billion people, developing abundant, affordable and clean energy, achieving light speed travel, etc. However, it is disastrous when dealing across the breath of human relations where self-sacrifice and accommodation are requisite to maintaining dynamic harmony/peace among human cultures, national governances, and biodiverse ecosystems.

As the community of AI professionals discusses the problem of control for future superintelligent agents, there are two often under-examined questions that linger in the background. **First, since humanity across the millennia has not been able to establish and abide by a universal convention of ethical reasoning, how can we expect AI professionals to solve this complex question as they write code for the first wave of artificial general intelligent agents?** More likely, governments/businesses will design artificial general intelligence (AGI) that will dominate other AGIs built by competing nations and businesses. Thus, a default ethic of power may reign supreme within future SIA as they emerge.

Second, since ethical reasoning among humans inevitably involves their various views of reality—their worldviews including religious/spiritual persuasions—how will AI professionals write code that will account for all secular and religious ethical reasoning? Will AI professionals employ a persuasion of reality to be agreed upon within one community—perhaps humanism that might create an artilect deity within multiple universes or a simulation universe created by tech savvy students who have mastered the craft of self-awareness and consciousness? Or will AI professionals incorporate religious worldviews in order for SIA to negotiate the nuances of human interactions. Science and religion have a tattered past thus making any attempt to account for religion within AGI an uncomfortable endeavor for the community of AI professionals. The purpose of this paper is to offer a pathway forward into the necessary ventures of ethics and religion.

In *Living with Superintelligent Agents: A Programmable Model for Ethical Controls of Future Artificial Superintelligence*, I posited a model for writing code to establish ethical reasoning for artificial general intelligences that would account for the diversity of human ethical reasoning that underlies specific ethical laws and general ethical rules of thumb. This ethical DNA model (eDNA) is based in 3-D overlapping geometries that are programmable. This model is supported by *Meta-language for Ethical Reasoning* (Ennis, 2015) which presents a means for establishing ethical reasoning across all cultures. The model accounts for logics of intellect and emotions as well as imagination and beauty.

Regarding religions: all religions are ways of seeing reality. Whether they conflict in their views or not, religions are subject to common reasoning on two interactive parameters: authoritative control across spatial perspectives. Both of these parameters can be mapped using the ethical DNA model. Authoritative control across spatial perspectives impacts all forms of religious/spiritual/worldview persuasions: monotheism, animism, polytheism, humanism, illusions, etc. As cultures focus their perspective of reality from a particular spatial point of reference (from self-awareness to transcendence) and assign a sense of authority (right of dominance) within that space, each religious/spiritual persuasion becomes reasonable from that point of reference (see Figure 1).



Religious/spiritual/worldview persuasions involve addressing spatial perspectives: *"What is beyond?"* That is, what do I see reflecting back at me—what is my primary point of reference (my authoritative control) for reality?

If there is **nothing beyond** my mind,
then I may believe I am a god or goddess or all is an illusion.

If **society** is my primary point of reference,
then I may be a humanist, atheist or agnostic.

If nature, the living **ecosystem**, is my primary reference point,
then I may believe in animism.

If the **universe** or multiple universes is all there is beyond me,
then I may believe in pantheism.

If there is a **transcendent beyond**,
then I may believe in polytheism, Satanism or monotheism.

Figure 1: Spatial perspectives and religious persuasions

The central construct of logic of intellect within the eDNA model is space and can therefore account for the spatial nature of religions. The construct of authoritative control can be mapped along the axes of power, bonding and meaning within the model.

Though such mapping of religious constructs may appear to be irrelevant within the community of AI professionals that often selects humanism as its point of reference, I suggest that controls for future SIA must account for a plurality of spatial points of reference—including humanism, polytheism, monotheism, and trinitarian persuasions. If this is not achieved, SIA will not master the nuances of intercultural relations that include all religious persuasions. And if SIA does not master these nuances, the ethical default of effective and efficient acquisition of power will yield an unprecedented power struggle between humans and superintelligent agents—with SIA the likely winner. Can we not program SIA to reason ethically with self-sacrifice and accommodation across cultures and religions as it re-creates itself?

Ralph C. Ennis
August 5, 2105