Stages of Ethical Development in Artificial General Intelligence Systems

Ben GOERTZEL and Stephan Vladimir BUGAJ

Novamente LLC and AGIRI Institute, Washington, DC

Abstract. A novel theory of the stages of ethical development in intelligent systems is proposed, incorporating prior related theories by Kohlberg and Gilligan, as well as Piaget's theory of cognitive development. This theory is then applied to the ethical development of integrative AGI systems that contain components carrying out simulation and uncertain inference – the key hypothesis being that effective integration of these components is central to the ascent of the AGI system up the ethical-stage hierarchy.

Keywords. Intelligent virtual agents, ethics, stages of moral development, ethics of care, ethics of justice, uncertain inference, simulation, theory of mind.

Introduction

Ethical judgment, in human beings, is acquired via a complex combination of genetic wiring, explicit instruction, and embodied mutual experience. Developmental theorists have identified various stages that young humans pass through on their way to achieving mature ethical faculties. Here we review some of these stage-based models with a view toward how they may be integrated and then transplanted to the domain of AGI ethics. We also correlate these stages with the five-imperatives view of ethics we have presented in [1], showing how earlier stages focus on three of the imperatives, with the other two imperatives rising to the fore as later stages are entered.

Perry's [2, 3] and especially Piaget's [4] theories of developmental stages form a foundation for our considerations here, but the essential original contribution is an identification of Kolhberg's [5, 6] and Gilligan's [7] complementary theories of ethical developmental stages with different components of contemporary AGI architectures. Kolhberg's stages, with their focus on abstract judgment, work nicely as a roadmap for the ethical development of logical inference engines; whereas Gilligan's stages, with their focus on empathy, are more pertinent as a model of the ethical development of internal-simulation-based AI. This leads to a notion of integrative ethical development as consisting of coordinated inferential and simulative ethical development. Along these lines, we present a novel theory of the stages of ethical development, incorporating Gilligan, Kohlberg and Piaget, and then apply this theory to AGI systems.

While many of the ideas discussed here would be meaningful more generally, for sake of concreteness we mainly restrict attention here to ethical development within AGI systems that contain and integrate components carrying out both uncertain logical inference and simulation. The Novamente Cognition Engine (or NCE, see [8, 9]) is one example of such an AGI architecture, but far from the only possible one. In the context of this sort of AGI architecture, we argue, integrative ethical development emerges as a consequence of overall coordination of ongoing development between these two components.

1. Stages of Cognitive Development

The best known theory of cognitive development is that of Jean Piaget. In a prior paper [10] we have presented a slightly modified form of the Piagetan developmental hierarchy, defined as follows:

Table 1. Modified Piagetan Developmental Stages

Stage	Example
Infantile	During this stage, the child learns about himself and his environment through motor and reflex actions. Thought derives from sensation and movement. Object permanence and the distinction between self and other are among the things learned.
Concrete	A rich usage of language emerges here, along with the usage of symbols to represent objects, the ability and propensity to think about things and events that aren't immediately present, and a robust but not totally flexible theory of mind.
Formal	Ability to think abstractly and to make rational judgements about concrete or observable phenomena. Logical reasoning and systematic experimentation.
Reflexive	Ability to modify one's own modes of thinking, reasoning, experimentation and self-perception at a fundamental level.

In that same paper we have defined a specific theory explaining how these stages manifest themselves in the development of AGI systems based on uncertain logical inference:

Table 2. Piagetan Developmental Stages for Uncertain Inference Based AGI Systems

Stage	Operational Aspect
Infantile	Able to recognize patterns in and conduct inferences about the world, but only using simplistic hard-wired (not experientially learned) inference control schema.
Concrete	Able to carry out more complex chains of reasoning regarding the world, via using inference control schemata that adapt behavior based on experience (reasoning about a given case in a manner similar to prior cases).
Formal	Able to carry out arbitrarily complex inferences (constrained only by computational resources) via including inference control as an explicit subject of abstract learning.
Reflexive	Capable of self-modification of internal structures.

Also relevant is William Perry's [2, 3] theory of the stages ("positions" in his writings) of intellectual and ethical development, which constitutes a model of iterative

refinement of approach in the developmental process of coming to intellectual and ethical maturity. These form an analytical tool for discerning the modality of belief of an intelligence by describing common cognitive approaches to handling the complexities of real world ethical considerations.

Table 3. Perry's Developmental Stages [with corresponding Piagetan Stages in brackets]

Stage	Substages
Dualism / Received Knowledge [Infantile]	– Basic duality ("All problems are solvable. I must learn the correct solutions.") – Full dualism ("There are different, contradictory solutions to many problems. I must learn the correct solutions, and ignore the incorrect ones")
Multiplicity [Concrete]	– Early multiplicity ("Some solutions are known, others aren't. I must learn how to find correct solutions.") – Late Multiplicity: cognitive dissonance regarding truth. ("Some problems are unsolvable, some are a matter of personal taste, therefore I must declare my own intellectual path.")
Relativism / Procedural Knowledge [Formal]	– Contextual Relativism ("I must learn to evaluate solutions within a context, and relative to supporting observation.") – Pre-Commitment ("I must evaluate solutions, then commit to a choice of solution.")
Commitment / Constructed Knowledge [Formal / Reflexive]	– Commitment ("I have chosen a solution.") – Challenges to Commitment ("I have seen unexpected implications of my commitment, and the responsibility I must take.") – Post-Commitment ("I must have an ongoing, nuanced relationship to the subject in which I evaluate each situation on a case-by-case basis with respects to its particulars rather than an ad-hoc application of unchallenged ideology.")

2. Stages of Development of the Ethics of Justice

Complementing generic theories of cognitive development such as Piaget's and Perry's, theorists have also proposed specific stages of moral and ethical development. The two most relevant theories in this domain are those of Kohlberg and Gilligan, which we will review here, both individually and in terms of their integration and application in the AGI context.

Lawrence Kohlberg's [5, 6] moral development model, called the "ethics of justice" by Gilligan, is based on a rational modality as the central vehicle for moral development. This model is based on an impartial regard for persons, proposing that ethical consideration must be given to all individual intelligences without a priori judgment (prejudice). Consideration is given for individual merit and preferences, and the goals of an ethical decision are equal treatment (in the general, not necessarily the particular) and reciprocity. Echoing Kant's [11] categorical imperative, the decisions considered most successful in this model are those which exhibit "reversibility", where a moral act within a particular situation is evaluated in terms of whether or not the act would be satisfactory even if particular persons were to switch roles within the situation. In other words, a situational, contextualized "do unto others as you would have them do unto you" criteria. The ethics of justice can be viewed as three stages (each of which has six substages, on which we will not elaborate here):

Table 4. Kohlberg's Stages of Development of the Ethics of Justics

Stage	Substages
Pre-Conventional	– Obedience and Punishment Orientation – Self-interest orientation
Conventional	– Interpersonal accord (conformity) orientation – Authority and social-order maintaining (law and order) orientation
Post-Conventional	– Social contract (human rights) orientation – Universal ethical principles (universal human rights) orientation

In Kohlberg's perspective, cognitive development level contributes to moral development, as moral understanding emerges from increased cognitive capability in the area of ethical decision making in a social context. Relatedly, Kohlberg also looks at stages of social perspective and their consequent interpersonal outlook. These are correlated to the stages of moral development, but also map onto Piagetian models of cognitive development (as pointed out e.g. by Gibbs [12], who presents a modification/interpretation of Kohlberg's ideas intended to align them more closely with Piaget's). Interpersonal outlook can be understood as rational understanding of the psychology of other persons (a theory of mind, with or without empathy). Stage one, emergent from the infantile cognitive stage, is entirely selfish as only self awareness has developed. As cognitive sophistication about ethical considerations increases, so do the moral and social perspective stages. Concrete and formal cognition bring about the first instrumental egoism, and then social relations and systems perspectives, and from formal and then reflexive thinking about ethics comes the post-conventional modalities of contractualism and universal mutual respect.

Table 5. Kohlberg's Stages of Development of Social Perspective and Interpersonal Morals

Stage of Social Perspective	Interpersonal Outlook
Blind egoism	No interpersonal perspective. Only self is considered.
Instrumental egoism	See that others have goals and perspectives, and either conform to or rebel against norms.
Social Relationships perspective	Able to see abstract normative systems
Social Systems perspective	Recognize positive and negative intentions
Contractual perspective	Recognize that contracts (mutually beneficial agreements of any kind) will allow intelligences to increase the welfare of both.
Universal principle of mutual respect	See how human fallibility and frailty are impacted by communication.

2.1. Uncertain Inference and the Ethics of Justice

Taking cue from the analysis given in [10] of Piagetan stages in uncertain inference based AGI systems, we may explore the manifestation of Kohlberg's stages

in AGI systems of this nature. Uncertain inference seems generally well-suited as an ethical learning system, due to the nuanced ethical environment of real world situations. An uncertain inference system, as defined in that previous paper, consists of four components: a content representation scheme (e.g. predicate logic, term logic, fuzzy logic); an uncertainty representation scheme (e.g. fuzzy truth values, probability values, probability intervals, imprecise probabilities, indefinite probabilities); a set of inference rules (e.g. those used in the NARS [13] or PLN [14] inference systems; and a set of inference control schemata (which in the Novamente CognitionEngine (NCE) are provided when PLN is integrated into the overall NCE framework.)

In general, an uncertain inference system may be viewed as a framework for dynamically updating a probabilistically weighted semantic network based on new incoming information and based on new conclusions derived via combining nodes and links in the network in appropriate, probabilistically grounded ways.

Probabilistic knowledge networks can model belief networks, imitative reinforcement learning based ethical pedagogy, and even simplistic moral maxims. In principle, they have the flexibility to deal with complex ethical decisions, including not only weighted "for the greater good" dichotomous decision making, but also the ability to develop moral decision networks which do not require that all situations be solved through resolution of a dichotomy.

When more than one person is being affected by an ethical decision, making a decision based on reducing two choices to a single decision can often lead to decisions of dubious ethics. However, a sufficiently complex uncertain inference network can represent alternate choices in which multiple actions are taken that have equal (or near equal) belief weight but have very different particulars – but because the decisions are applied in different contexts (to different groups of individuals) they are morally equivalent.

Infantile and concrete cognition are the underpinnings of the egoist and socialized stages, with formal aspects also playing a role in a more complete understanding of social models when thinking using the social modalities. Cognitively infantile patterns can produce no more than blind egoism or compassion as without a theory of mind or a refined capability for empathy, there is no capability to consider the other in a contextually appropriate way. Since most intelligences acquire concrete modality and therefore some nascent social perspective relatively quickly, most egoists are instrumental egoists. The social relationship and systems perspectives include formal aspects which are achieved by systematic social experimentation, and therefore experiential reinforcement learning of correct and incorrect social modalities. Initially this is a one-on-one approach (relationship stage), but as more knowledge of social action and consequences is acquired, a formal thinker can understand not just consequentiality but also intentionality in social action.

Extrapolation from models of individual interaction to general social theoretic notions is also a formal action. Rational, logical positivist approaches to social and political ideas, however, are the norm of formal thinking. Contractual and committed moral ethics emerges from a higher-order formalization of the social relationships and systems patterns of thinking. Generalizations of social observation become, through formal analysis, systems of social and political doctrine. Highly committed, but grounded and logically supportable, belief is the hallmark of formal cognition as expressed in the contractual moral stage. Though formalism is at work in the socialized moral stages, its fullest expression is in committed contractualism.

Finally, reflexive cognition is especially important in truly reaching the post-commitment moral stage in which nuance and complexity are accommodated. Because reflexive cognition is necessary to change one's mind not just about particular rational ideas, but whole *ways of thinking*, this is a cognitive precedent to being able to reconsider an entire belief system, one that has had contractual logic built atop reflexive adherence that began in early development. If the initial moral system is viewed as positive and stable, than this cognitive capacity is seen as dangerous and scary, but if early morality is stunted or warped, then this ability is seen as enlightened. However, achieving this cognitive stage does not mean one automatically changes their belief systems, but rather that the mental machinery is in place to consider the possibilities. Because many people do not reach this level of cognitive development in the area of moral and ethical thinking, it is associated with negative traits ("moral relativism" and "flip-flopping"). However, this cognitive flexibility generally leads to more sophisticated and applicable moral codes, which in turn leads to morality which is actually more stable because it is built upon extensive and deep consideration rather than simple adherence to reflexive or rationalized ideologies.

3. Stages of Development of Empathic Ethics

Complementing Kohlberg's logic-and-justice-focused approach, Carol Gilligan's [7] "ethics of care" model is a moral development theory which posits that empathetic understanding plays the central role in moral progression from an initial self-centered modality to a socially responsible one. For this approach to be applied in an AGI, the AGI must be capable of internal simulation of other minds it encounters, in a similar manner to how humans regularly simulate one another internally [15]. Without any mechanism for internal simulation, it is unlikely that an AGI can develop any sort of empathy toward other minds, as opposed to merely logically or probabilistically modeling other agents' behavior or other minds' internal contents.

The ethics of care model is concerned with the ways in which an individual cares (responds to dilemmas using empathetic responses) about self and others. The ethics of care is broken into the same three primary stage as Kohlberg, but with a focus on empathetic, emotional caring rather than rationalized, logical principles of justice:

Table 6. Gilligan's Stages of the Ethics of Care

Stage	Principle of Care
Pre-Conventional	Individual Survival
Conventional	Self Sacrifice for the Greater Good
Post-Conventional	Principle of Nonviolence (do not hurt others, or oneself)

In Gilligan's perspective, the earliest stage of ethical development occurs before empathy becomes a consistent and powerful force. Next, the hallmark of the conventional stage is that at this point, the individual is so overwhelmed with their empathic response to others that they neglect themselves in order to avoid hurting others. Note that this stage doesn't occur in Kohlberg's hierarchy at all. Kohlberg and Gilligan both begin with selfish unethicality, but their following stages diverge. A person could in principle manifest Gilligan's conventional stage without having a refined sense of justice (thus not entering Kohlberg's conventional stage); or they could manifest Kohlberg's conventional stage without partaking in an excessive degree of self-sacrifice (thus not entering Gilligan's conventional stage). We will suggest below

that in fact the empathic and logical aspects of ethics are more unified in real human development than these separate theories would suggest.

It is interesting to note that Gilligan's and Kohlberg's final stages converge more closely than their intermediate ones. Kohlberg's post-conventional stage focuses on universal rights, and Gilligan's on universal compassion. Still, the foci here are quite different; and, as will be elaborated below, we believe that both Kohlberg's and Gilligan's theories constitute very partial views of the actual end-state of ethical advancement.

Gilligan's theory was proposed partly as a reaction to the perceived male bias of Kohlberg's thoery. There is certainly some meat to this complaint, as there is much evidence that females tend to be more empathic in their response to ethical judgment, whereas men tend to be more focused on abstract notions of rights and fairness. In general, however, we feel that, just as Kohlberg gives short shrift to empathy, Gilligan gives short shrift to logical reasoning, and that due to these limitations of perspective, both theorists have failed to create adequately scoped theories of ethical development.

4. An Integrative Approach to Ethical Development

We deny the false dichotomy of a "feminine" ethics of care vs. a "masculine" ethics of justice, and propose that both Kohlberg's and Gilligan's theories contain elements of the whole picture of ethical development, and that both approaches are necessary to create a moral, ethical artificial general intelligence -- just as, we suggest, both internal simulation and uncertain inference are necessary to create a sufficiently intelligent and volitional intelligence in the first place. Also, we contend, the lack of direct analysis of the underlying psychology of the stages is a deficiency shared by both the Kohlberg and Gilligan models as they are generally discussed. A successful model of integrative ethics necessarily contains elements of both the care and justice models, as well as reference to the underlying developmental psychology and its influence on the character of the ethical stage.

With these notions in mind, we propose the following integrative theory of the stages of ethical development, shown in the table at the end of this section.

In our integrative model, the justice-based and empathic aspects of ethical judgment are proposed to develop together. Of course, in any one individual, one or another aspect may be dominant. Even so, however, the combination of the two is equally important as either of the two individual ingredients.

For instance, we suggest that in any psychologically healthy human, the conventional stage of ethics (typifying childhood, and in many cases adulthood as well) involves a combination of Gilligan-esqe empathic ethics and Kohlberg-esque ethical reasoning. This combination is supported by Piagetan concrete operational cognition, which allows moderately sophisticated linguistic interaction, theory of mind, and symbolic modeling of the world. And, similarly, we propose that in any truly ethically mature human, empathy and rational justice are both fully developed. Indeed the two interpenetrate each other deeply.

Once one goes beyond simplistic, childlike notions of fairness ("an eye for an eye" and so forth), applying rational justice in a purely intellectual sense is just as difficult as any other real-world logical inference problem. Ethical quandaries and quagmires are easily encountered, and are frequently cut through by a judicious application of empathic simulation.

On the other hand, empathy is a far more powerful force when used in conjunction with reason: analogical reasoning lets us empathize with situations we have never experience. For instance, a person who has never been clinically depressed may have a hard time empathizing with individuals who are; but using the power of reason, they can imagine their worst state of depression magnified by several times and then extended over a long period of time, and then reason about what this might be like ... and empathize based on their inferential conclusion. Reason is not antithetical to empathy but rather is the key to making empathy more broadly impactful.

Finally, the enlightened stage of ethical development involves both a deeper compassion and a more deeply penetrating rationality and objectiveness. Empathy with all sentient beings is manageable in everyday life only once one has deeply reflected on one's own self and largely freed oneself of the confusions and illusions that characterize much of the ordinary human's inner existence. It is noteworthy, for example, that Buddhism contains both a richly developed ethics of universal compassion, and also an intricate logical theory of the inner workings of cognition [16], detailing in exquisite rational detail the manner in which minds originate structures and dynamics allowing them to comprehend themselves and the world.

4.1. The Stages of Ethical Development and the Five Ethical Imperatives

In [1] we have proposed a series of five ethical imperatives and explored their implications for interacting with, and teaching ethics to, AGI systems: 1. the *imitability imperative* (i.e. the folk Golden Rule "do unto others as one would have them do unto you") fairly narrowly and directly construed): the goal of acting in a way so that having others directly imitate one's actions, in directly comparable contexts, is desirable to oneself; 2. the *comprehensibility imperative*: the goal of acting in a way so that others can understand the principles underlying one's actions; 3. *experiential groundedness*. An intelligent agent should not be expected to act according to an ethical principle unless there are many examples of the principle-in-action in its own direct or observational experience; 4. Kant's *categorical imperative*: choose behaviors according to a certain maxim only if you would will that maxim to be followed universally by all sentient beings; 5. *logical coherence.* An ethical system should be roughly logically coherent, in the sense that the different principles within it should mesh well with one another and perhaps even naturally emerge from each other.

Specific ethical qualities corresponding to the five imperatives have been italicized in the above table of developmental stages. Firstly, it seems that imperatives 1-3 are critical for the passage from the pre-ethical to the conventional stages of ethics. A child learns ethics largely by copying others, and by being interacted with according to simply comprehensible implementations of the Golden Rule. In general, when interacting with children learning ethics, it is important to act according to principles they can comprehend. And given the nature of the concrete stage of cognitive development, experiential groundedness is a must.

As a hypothesis regarding the dynamics underlying the psychological development of conventional ethics, what we propose is as follows: The emergence of concrete-stage cognitive capabilities leads to the capability for fulfillment of ethical imperatives 1 and 2 – a comprehensible and workable implementation of the Golden Rule, based on a

Table 7. Integrative Model of the Stages of Ethical Development

Stage	Characteristics
Pre-ethical	– Piagetan infantile to early concrete (aka pre-operational) – Radical selfishness or selflessness may, but do not necessarily, occur – No coherent, consistent pattern of consideration for the rights, intentions or feelings of others – Empathy is generally present, but erratically
Conventional Ethics	– Concrete cognitive basis – Perry's Dualist and Multiple stages – *The common sense of the Golden Rule is appreciated, with cultural conventions for abstracting principles from behaviors* – *One's own ethical behavior is explicitly compared to that of others* – Development of a functional, though limited, theory of mind – Ability to intuitively conceive of notions of fairness and rights – Appreciation of the concept of law and order, which may sometimes manifest itself as systematic obedience or systematic disobedience – Empathy is more consistently present, especially with others who are directly similar to oneself or in situations similar to those one has directly experienced – Degrees of selflessness or selfishness develop based on ethical groundings and social interactions.
Mature Ethics	– Formal cognitive basis – Perry's Relativist and "Constructed Knowledge" stages – *The abstraction involved with applying the Golden Rule in practice is more fully understood and manipulated, leading to limited but nonzero deployment of the Categorical Imperative* – *Attention is paid to shaping one's ethical principles into a coherent logical system* – Rationalized, moderated selfishness or selflessness. – Empathy is extended, using reason, to individuals and situations not directly matching one's own experience – Theory of mind is extended, using reason, to counterintuitive or experientially unfamiliar situations – Reason is used to control the impact of empathy on behavior (i.e. rational judgments are made regarding when to listen to empathy and when not to) – Rational experimentation and correction of theoretical models of ethical behavior, and reconciliation with observed behavior during interaction with others. – Conflict between pragmatism of social contract orientation and idealism of universal ethical principles. – Understanding of ethical quandaries and nuances develop (pragmatist modality), or are rejected (idealist modality). – Pragmatically critical social citizen. Attempts to maintain a balanced social outlook. Considers the common good, including oneself as part of the commons, and acts in what seems to be the most beneficial and practical manner.
Enlightened Ethics	– Reflexive cognitive basis – *Permeation of the categorical imperative and the quest for coherence through inner as well as outer life* – Experientially grounded and logically supported rejection of the illusion of moral certainty in favor of a case-specific analytical and empathetic approach that embraces the uncertainty of real social life – Deep understanding of the illusory and biased nature of the individual self, leading to humility regarding one's own ethical intuitions and prescriptions – Openness to modifying one's deepest, ethical (and other) beliefs based on experience, reason and/or empathic communion with others – Adaptive, insightful approach to civil disobedience, considering laws and social customs in a broader ethical and pragmatic context – Broad compassion for and empathy with all sentient beings

	– A recognition of inability to operate at this level at all times in all things, and a vigilance about self-monitoring for regressive behavior.

combination of inferential and simulative cognition (operating largely separately at this stage, as will be conjectured below). The effective interoperation of ethical imperatives 1-3, enacted in an appropriate social environment, then leads to the other characteristics of the conventional ethical stage. The first three imperatives can thus be viewed as the seed from which springs the general nature of conventional ethics.

On the other hand, logical coherence and the categorical imperative (imperatives 4 and 5) are matters for the formal stage of cognitive development, which come along only with the mature approach to ethics. These come from abstracting ethics beyond direct experience and manipulating them abstractly and formally – a stage which has the potential for more deeply and broadly ethical behavior, but also for more complicated ethical perversions (it is the mature capability for formal ethical reasoning that is able to produce ungrounded abstractions such as "I'm torturing you for your own good"). Developmentally, we suggest that once the capability for formal reasoning matures, the categorical imperative and the quest for logical ethical coherence naturally emerge, and the sophisticated combination of inferential and simulative cognition embodied in an appropriate social context then result in the emergence of the various characteristics typifying the mature ethical stage.

Finally, it seems that one key aspect of the passage from the mature to the enlightened stage of ethics is the penetration of these two final imperatives more and more deeply into the judging mind itself. The reflexive stage of cognitive development is in part about seeking a deep logical coherence between the aspects of one's own mind, and making reasoned modifications to one's mind so as to improve the level of coherence. And, much of the process of mental discipline and purification that comes with the passage to enlightened ethics has to do with the application of the categorical imperative to one's own thoughts and feelings – i.e. making a true inner systematic effort to think and feel only those things one judges are actually generally good and right to be thinking and feeling. Applying these principles internally appears critical to effectively applying them externally, for reasons that are doubtless bound up with the interpenetration of internal and external reality within the thinking mind, and the "distributed cognition" phenomenon wherein individual mind is itself an approximative abstraction to the reality in which each individual's mind is pragmatically extended across their social group and their environment [17].

5. Integrative Ethics and Integrative Artificial General Intelligence

And what does our integrative approach to ethical development have to say about the ethical development of AGI systems? The lessons are relatively straighforward, if one considers an AGI system that, like the Novamente Cognition Engine (NCE), explicitly contains components dedicated to logical inference and to simulation. Application of the above ethical ideas to other sorts of AGI systems is also quite possible, but would require a lengthier treatment and so won't be addressed here.

In the context of a NCE-type AGI system, Kolhberg's stages correspond to increasingly sophisticated application of logical inference to matters of rights and fairness. It is not clear whether humans contain an innate sense of fairness. In the context of AGIs, it would be possible to explicitly wire a sense of fairness into an AGI system, but in the context of a rich environment and active human teachers, this

actually appears quite unnecessary. Experiential instruction in the notions of rights and fairness should suffice to teach an inference-based AGI system how to manipulate these concepts, analogously to teaching the same AGI system how to manipulate number, mass and other such quantities. Ascending the Kohlberg stages is then mainly a matter of acquiring the ability to carry out suitably complex inferences in the domain of rights and fairness. The hard part here is inference control – choosing which inference steps to take – and in a sophisticated AGI inference engine, inference control will be guided by experience, so that the more ethical judgments the system has executed and witnessed, the better it will become at making new ones. And, as argued above, simulative activity can be extremely valuable for aiding with inference control. When a logical inference process reaches a point of acute uncertainty (the backward or forward chaining inference tree can't decide which expansion step to take), it can run a simulation to cut through the confusion – i.e., it can use empathy to decide which logical inference step to take in thinking about applying the notions of rights and fairness to a given situation.

Gilligan's stages correspond to increasingly sophisticated control of empathic simulation – which in a NCE-type AGI system, is carried out by a specific system component devoted to running internal simulations of aspects of the outside world, which includes a subcomponent specifically tuned for simulating sentient actors. The conventional stage has to do with the raw, uncontrolled capability for such simulation; and the post-conventional stage corresponds to its contextual, goal-oriented control. But controlling empathy, clearly, requires subtle management of various uncertain contextual factors, which is exactly what uncertain logical inference is good at – so, in an AGI system combining an uncertain inference component with a simulative component, it is the inference component that would enable the nuanced control of empathy allowing the ascent to Gilligan's post-conventional stage.

In our integrative perspective, in the context of an AGI system integrating inference and simulation components, we suggest that the ascent from the pre-ethical to the conventional stage may be carried out largely via independent activity of these two components. Empathy is needed, and reasoning about fairness and rights are needed, but the two need not intimately and sensitively intersect – though they must of course intersect to some extent.

The main engine of advancement from the conventional to mature stage, we suggest, is robust and subtle integration of the simulative and inferential components. To expand empathy beyond the most obvious cases, analogical inference is needed; and to carry out complex inferences about justice, empathy-guided inference-control is needed.

Finally, to advance from the mature to the enlightened stage, what is required is a very advanced capability for unified reflexive inference and simulation. The system must be able to understand itself deeply, via modeling itself both simulatively and inferentially – which will generally be achieved via a combination of being good at modeling, and becoming less convoluted and more coherent, hence making self-modeling eacher.

Of course, none of this tells you in detail how to create an AGI system with advanced ethical capabilities. What it does tell you, however, is one possible path that may be followed to achieve this end goal. If one creates an integrative AGI system with appropriately interconnected inferential and simulative components, and treats it compassionately and fairly, and provides it extensive, experientially grounded ethical instruction in a rich social environment, then the AGI system should be able to ascend

the ethical hierarchy and achieve a high level of ethical sophistication. In fact it should be able to do so more reliably than human beings because of the capability we have to identify its errors via inspecting its internal knowedge-stage, which will enable us to tailor its environment and instructions more suitably than can be done in the human case.

If an absolute guarantee of the ethical soundness of an AGI is what one is after, the line of thinking proposed here is not at all useful. However, if what one is after is a plausible, pragmatic path to architecting and educating ethical AGI systems, we believe the ideas presented here constitute a sensible starting-point. Certainly there is a great deal more to be learned and understood – the science and practice of AGI ethics, like AGI itself, are at a formative stage at present. What is key, in our view, is that as AGI technology develops, AGI ethics develops alongside and within it, in a thoroughly coupled way.

References

[1] Bugaj, Stephan Vladimir and Ben Goertzel (2007). Five Ethical Imperatives for AGI Systems. This volume.

[2] Perry, William G., Jr. Forms of Intellectual and Ethical Development in the College Years: A Scheme. New York: Holt, Rinehart and Winston, 1970.

[3] Perry, William G., Jr. "Cognitive and Ethical Growth: The Making of Meaning", in Arthur W. Chickering and Associates, The Modern American College, San Francisco: Jossey-Bass pp 76-116. 1981.

[4] Piaget, Jean. “Piaget’s theory.” In P. Mussen (ed). Handbook of Child Psychology. 4th edition. Vol. 1. New York: Wiley, 1983.

[5] Kohlberg, Lawrence; Charles Levine, Alexandra Hewer (1983). Moral stages : a current formulation and a response to critics. Basel, NY: Karger.

[6] Kohlberg, Lawrence. Essays on Moral Development, Vol. I: The Philosophy of Moral Development. Harper & Row, 1981.

[7] Gilligan, Carol (1982). In a Different Voice. Cambridge, MA: Harvard University Press, 1982.

[8] Goertzel, Ben (2006). The Hidden Pattern. BrownWalker Press

[9] Goertzel, Ben, Moshe Looks and Cassio Pennachin (2004). Novamente: An Integrative Architecture for Artificial General Intelligence. Proceedings of AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, Washington DC, August 2004

[10] Goertzel, Ben and Stephan Vladimir Bugaj (2006). Stages of Cognitive Development in Uncertain-Logic-Based AGI Systems. In Advances in artificial general intelligence, Ed. by Ben Goertzel and Pei Wang:36-54. Amsterdam: IOS Press.

[11] Kant, Immanuel (1964). Groundwork of the Metaphysic of Morals. Harper and Row Publishers, Inc.

[12] Gibbs, John (1978). “Kohlberg’s moral stage theory: a Piagetian revision.” Human Development, 1978, 22, 89-112

[13] Wang, Pei (2006). Rigid Flexibility: The Logic of Intelligence. Springer-Verlag

[14] Ikle, Matthew and Ben Goertzel (2006). Indefinite Probabilities for General Intelligence. In Advances in Artificial General Intelligence, Edited by Ben Goertzel and Pei Wang, IOS Press

[15] Gordon, Robert (1986). Folk Psychology as Simulation. Mind and Language, 1, 158-171,

[16] Stcherbatsky, Theodore (2000). Buddhist Logic. Motilal Banarsidass Pub, New York

[17] Hutchins, E. (1995) Cognition in the Wild (ISBN 0-262-58146-9