Ethics as Aesthetic: A Computational Creativity Approach to Ethical Behavior

Dan Ventura[1] and **Darin Gates**[1,2,3]
[1]Computer Science Department, [2]Department of Philosophy, [3]The Wheatley Institution
Brigham Young University
ventura@cs.byu.edu, gatesdarin@gmail.com

Abstract

We address the question of how to build AI agents that behave ethically by appealing to a computational creativity framework in which output artifacts are agent behaviors and candidate behaviors are evaluated using a normative ethics as the aesthetic measure. We then appeal again to computational creativity to address the meta-level question of which normative ethics the system should employ as its aesthetic, where now output meta-artifacts are normative ethics and candidate ethics are evaluated using a meta-level-ethics-based aesthetic. We consider briefly some of the issues raised by such a proposal as well as how the hybrid base-meta-level system might be evaluated from three different perspectives: creative, behavioral and ethical.

Introduction

Artificial intelligence (AI) continues to mature and deliver on promises 50 years or more in the making, and this development has been especially marked in the last decade. However, as significant as these AI advances have become, the ultimate goal of artificial general intelligence is yet to be realized. Nevertheless, a great deal has been said about ethical issues arising from the development of AI systems (both the current specialized variety and the yet-quixotic general variety) that now can or may soon be able to impact humanity at unprecedented scale, with predictions ranging from the possibility of a Utopian post-human immortality to the enslavement or even annihilation of the human race. Such discussions appear in every form imaginable, from monographs (Wallach and Allen 2008; Anderson and Leigh 2011; Müller 2016) to academic journals (Anderson and Anderson 2006; Muehlhauser and Helm 2012) to popular literature (Kurzweil 2005; McGee 2007; Fox 2009; Coeckelbergh 2014) to government studies (Lin, Bekey, and Abney 2008; European Parliament, Committee on Legal Affairs 2017). These treatments almost always take the form of applied ethics, either to be applied to humans doing the research that will inevitably lead to an AI-dominated future or to be applied to the AI systems themselves, or both. These discussions are most often normative in nature. Thus, we currently face the twin problems:

1. How can we ensure an AI agent behaves ethically?
2. What do we mean by ethical?

To begin with, we will simply postulate an abstract computational creativity (CC) approach for the implementation of an AI system. That is, we postulate a system whose domain of creation is behavioral policy, a system whose output artifacts are goals and/or decisions and/or sequences of actions. Given this admittedly ambitious premise and using a CC framework, we will argue the two questions can be naturally addressed. The question of how to impose an ethics on such a system can be addressed by implementing the CC system's aesthetic for evaluating artifacts as a (normative) ethics. In other words, that ethics acts as the filter by which the utility of system actions, decisions and goals is judged. The meta-level question of *which* normative ethics ought to be applied as the system's aesthetic can be addressed by allowing the system to create a suitable norm, given some meta-level aesthetic for ethics. That is, we suggest a CC system whose output artifact is a normative ethics and whose aesthetic is some way to evaluate said norm.

To summarize, we propose an appeal to computational creativity that answers both of our questions of interest:

1. We can build an ethical AI agent as a computational creativity system whose output artifacts are goals, decisions and behaviors and whose aesthetic component is a normative ethics.
2. We can delegate the choice of normative ethics to the AI agent by implementing a meta-level computational creativity system whose output artifacts are normative ethics and whose aesthetic is a meta-level ethics.

Ethical Behavior Invention

The field of computational creativity has been described as "the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative" (Colton and Wiggins 2012). It has been characterized by attempts at building systems for meeting this standard in a wide variety of domains, including culinary recipes (Morris et al. 2012; Varshney et al. 2013), language constructs such as metaphor (Veale and Hao 2007) and neologism (Smith, Hintze, and Ventura 2014), visual art (Colton 2012; Norton, Heath, and Ventura 2013), poetry (Toivanen et al. 2012; Oliveira 2012; Veale 2013), humor (Binsted and Ritchie 1994; Stock and Strapparava

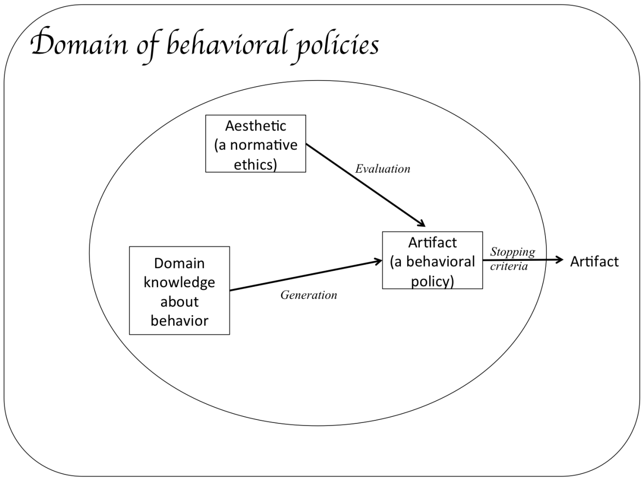
Domain of behavioral policies
Aesthetic (a normative ethics)
Evaluation
Domain knowledge about behavior
Generation
Artifact (a behavioral policy)
Stopping criteria
Artifact

Figure 1: A CC system embedded in the domain of behavioral policies uses domain knowledge about behavior to generate candidate policies that are vetted by an ethics-based aesthetic. Those polices judged to be of value by the aesthetic are exported to the domain, becoming viable policies for an AI agent.

2003), advertising and slogans (Strapparava, Valitutti, and Stock 2007; Özbal, Pighin, and Strapparava 2013), narrative and story telling (Pérez y Pérez and Sharples 2004; Riedl and Young 2010), mathematics (Colton, Bundy, and Walsh 1999), games (Liapis, Yannakakis, and Togelius 2012; Cook, Colton, and Gow 2016) and music (Bickerman et al. 2010; Pachet and Roy 2014).

Recently an abstract approach to building such a system for *any* domain has been proposed (Ventura 2017), with the goal being an autonomous CC system that intentionally produces artifacts that are both novel and valuable in a particular domain. The system has a domain-specific *knowledge base*; it has a domain-appropriate *aesthetic*; and it has the ability to externalize artifacts that potentially can contribute to the domain. The system incorporates additional components as well, but they will not be important for the current discussion and the reader is referred to the original paper for more details.

We consider an AI agent as a CC system whose domain of creation is behavioral policy, and a simple abstraction of this idea is shown in Fig 1. The system creates behavior policies by generated candidate policies based on its domain knowledge, and it evaluates those candidate policies using an aesthetic that is a normative ethics. For example, suppose the system incorporates a simple hedonistic ethics that values knowledge acquisition as its aesthetic and that it generating the candidate behaviors *read Wikipedia* and *find charging station*. The former goal will be evaluated more favorably than the latter and may be output as a viable output artifact if that evaluation is above a threshold. Or, suppose the system's aesthetic is implemented as a Kantian-style ethics focused on the duty of delivering its payload and that it generates the same two candidate behaviors. Now, neither may be evaluated very favorably and both might be discarded; however, if the agent's power level is too low to allow completion of a delivery, the latter may instead be selected as a high-quality behavior.

Given this framework, we can argue that, assuming an appropriate ethics, the system will behave ethically—it will not produce any actions that do not meet some ethical threshold and are thus judged of high-enough value to be output as viable. As an obvious example of being above a certain threshold, an AI agent would not deliver its payload if that would involve harming someone—a clear example of violating Kant's principle that we ought not to treat someone merely as a means (1994)—or if perhaps it determined that delivering its payload would prevent another important obligation. Thus, the threshold would be something like *help fulfill the duty to deliver a payload, effectively and on time, unless doing so would seriously harm another person, etc.* Here we of course run into the problem of prioritization in the face of conflicting duties. We will say more about this issue shortly, but we will at least note here that it may be desirable for an AI agent to have the ability to act in ways that are analogous to the types of special obligations we have as humans (while at the same time also allowing creative behaviors within certain ethical boundary constraints). For example, perhaps a domestic companion robot would give significantly higher weight to the needs of the person to whom it is assigned: helping its companion would take priority over the possibility of helping others. However, we could also allow for the possibility of the robot to decide to not help the assigned companion in certain emergency cases in which another person nearby needed life-saving attention, just as we would expect a parent to prioritize helping a stranger in serious need over the needs of his or her child in certain cases (i.e., as long as the need of the child is minor).

This leaves us with two challenges: what is an appropriate ethics and how can it be operationalized? The first of these is, of course, a fundamental question that is thousands of years old. The second is much more recent and has likely only become significant in the past 50 years. Both questions are beyond the scope of this treatment, but it is likely the case that there is no single answer to the former question, at least with respect to AI systems,[1] as most famously demonstrated by Asimov's examination of his *Three Laws of Robotics* (1950). It is also very possibly the case that a satisfactory answer to the second question requires and/or will result in a greater understanding of human ethics. And, just as in the case of an examination of human ethics, these questions somewhat naturally lead us to meta-level ethical questions.

By what principles should our CC system be governed? One attractive possibility is the adoption of an utilitarian-consequentialist ethics, due to the conceptual simplicity of choosing the action that maximizes the overall-good (or at least brings about the most utility for all those who are concerned). However, such utilitarian-consequentialism faces the serious objection that it would permit widespread violation of constraints against harm doing in the name of such supposed optimization. Examples such as the well-

[1] And likely with respect to humans as well, actually.

known Transplant scenario illustrate this concern (Kagan 1998). In this scenario, a surgeon would involuntarily sacrifice one innocent person to use his organs to save five others. Such actions clearly violate serious *negative duties* (duties not to harm) for the sake of *positive duties* (duties to help). While there may well be certain thresholds at which even non-consequentialists would agree that some such decision would be justifiable, perhaps most people would argue that there should be near absolute constraints against such actions. For utilitarian-consequentialism, all that matters is that the overall harm is minimized. It does not matter whether negative duties (duties of non-harm) are violated to minimize harm/maximize utility. However, for non-consequentialists (deontologists), there is an asymmetry between negative and positive duties. Negative duties are much more stringent in the sense that their violation requires an overwhelming amount of good (or harm prevention) to be justified. Common sense morality is most likely more in line with such non-consequentialist intuitions.

When it comes to many everyday ethical decisions, there is, of course, significant agreement between the major ethical approaches: utilitarian-consequentialism, non-consequentialism (e.g. Kantian ethics) and virtue ethics. Their divergence becomes obvious only in extreme situations in which maximizing the overall good violates the most serious ethical constraints—constraints against harming innocent people, privacy violation, and so on. Cases such as these are obviously relevant to unavoidable harm scenarios, such as those faced by self-driving vehicles.[2]

One ethical framework that might offer a helpful model for an AI agent ethics is intuitionism.[3] Intuitionism holds that there are categories of *prima facie* duties that are *self-evident*, *non-absolute*, and always *morally relevant*. These duties are **non-injury** (non-maleficence), **beneficence**, **veracity**, **fidelity**, **gratitude**, **justice**, **self-improvement**, and **reparation**.[4] For intuitionism, while it is self-evident that we are morally constrained by these *prima facie* duties, it is not always self-evident what is the right action in situations in which there are conflicting duties at play. Intuitionism gives us neither a weighted hierarchy, nor a decision procedure for how to choose the right action in such conflicts. Instead, it assumes we will need to make a reasoned judgment to decide which duty (or duties) deserves more weight in a particular situation. However, it may be possible to come up with factors that help make such decisions, and a CC AI agent might be capable of so doing.[5] Intuitionism thus offers some of the flexibility people find attractive in utilitarianism, while at the same time offering important constraints against the worst implications that certainly seem to follow from a straightforward use of the utilitarian maximization principle.

When human agents decide in favor of one moral rule/principle over another (in such conflict situations), we assume there should be a plausible account of why such a decision was made. This is not to suggest that we expect said person to have pre-emptively produced such a justification, nor even that they ever explicitly work-out an account of why they acted as they did—though in cases where there was sufficient time for deliberation the person may, indeed, have thought through such an account. However, we expect that such justification *is* possible, at least *post hoc*. Similarly, we are interested in whether it is possible for an AI agent to develop something like good ethical *judgment* (that can therefore be justified).

The hope is that such an AI agent could find ways to produce ethical decisions that would be plausible (given certain constraints) and yet also be surprising in the way that they solve ethical quandaries, without the necessity of a fully worked-out super ethics. In other words, we are suggesting a solution to what Bostrom calls the "ultimate challenge of machine ethics"—namely, "How do you build an AI which behaves more ethically than you?" As he writes:

> This is not like asking our own philosophers to produce superethics, any more than Deep Blue was constructed by getting the best human chess players to program in good moves (Bostrom and Yudkowski 2014).

If we build into our AI agents governing principles (including serious constraints on harm doing) that attempt to mirror those common and significant ethical principles shared by the major schools of thought, we will be more likely to end up with actions that most people would consider ethical. Thus, just as we can characterize a successful CC system as one that exhibits behaviors that unbiased observers would deem to be creative, so we could describe a successful CC system for inventing ethical behavior as one that *behaves such that an unbiased observer would deem it to be ethical*. And, just as computational creative agents will create in ways that surprise but yet are in harmony with certain generally determined (domain-specific) principles, so the hope is that an ethical CC system/agent would, similarly, be ethical in ways that would surprise us, and yet still be in harmony with what an unbiased observer would agree is ethically acceptable.

One way to formulate the goal of a CC ethics would be the production of ethical decisions untainted by human biases and rationalizations while utilizing the quality of judgment, sensitivity, and wisdom that we as humans exercise (at our best) when deciding between conflicting duties.

Normative Ethics Invention

If we can postulate a CC system that creates behaviors and evaluates their aesthetic value via some ethics, why not pos-

[2] Arguably, these scenarios will be rare. Arguably too, we should not hold the development of self-driving cars hostage to these possibilities. As has been often pointed out, around 94% of serious injuries/fatalities that occur in car accidents come from human error, which would be greatly reduced were widespread implementation of self-driving cars to become a reality.

[3] Intuitionism was formulated by the 20th century Oxford moral philosopher W.D. Ross.

[4] W.D. Ross (the founder of intuitionism) originally postulated seven categories of prima facie duties. Here we follow Robert Audi's addition of veracity (which for Ross was implied in fidelity) (Audi 2009).

[5] Such factors might include the type of special obligations we mentioned earlier, as well as other factors such as the *magnitude of consequences*, the *probability of effect*, *temporal immediacy*, *proximity*, *concentration of the effect*, and so on (Jones 1991).

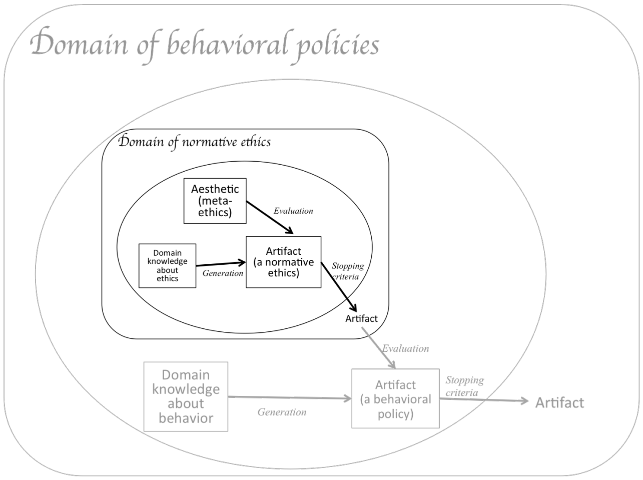
Domain of behavioral policies
Domain of normative ethics
Aesthetic (meta-ethics)
Evaluation
Domain knowledge about ethics
Generation
Artifact (a normative ethics)
Stopping criteria
Artifact
Evaluation
Domain knowledge about behavior
Generation
Artifact (a behavioral policy)
Stopping criteria
Artifact

Figure 2: A meta-level CC system for creating normative ethics whose output artifact (a normative ethics) is used as the aesthetic in the base-level system of Fig. 1.

tulate a meta-level CC system that creates normative ethics and evaluates their meta-aesthetic value using some meta-level ethics? This system naturally solves both of the outstanding questions above.[6] Fig. 2 shows how this meta-level system is incorporated into the base-level system of Fig. 1. The base-level, behavioral system appeals to the meta-level, ethical system to create a "good" normative ethics that it then uses as its aesthetic to judge candidate actions. For example, the meta-level ethics might require a well-formed semantics and justifiability, and candidate normative ethics that can be shown to have both of these qualities would be evaluated as (meta-)aesthetically valuable, while those that possess one of the qualities would be evaluated as less valuable.

We are again in a position to argue that, assuming an appropriate meta-level ethics, the (base-level) system will behave ethically—it will still not produce any actions that do not meet some ethical threshold and are thus judged of high-enough value to be output as valuable (in an ethical sense). Notably, this argument now does not depend on the assumption of an appropriate ethics—we have eliminated this dependency by appealing to the meta-level. However, of course, we now have an assumption of an appropriate meta-level ethics, which immediately leads us back to the same difficult questions applied this time to the meta-level: what is an appropriate *meta-level* ethics and can *it* be operationalized? While we do not here offer a solution to either of these conundrums, it is possible that the more abstract nature of a meta-level ethics might admit fewer viable possibilities and thus afford us a chance as a field for coming to an agreement regarding the first problem. On the other hand, it is also possible that this additional abstraction may have just the opposite effect for the second problem, introducing additional difficulty in the operationalization of this agreed upon meta-level ethics.

Assuming we do find suitable answers to both of these meta-problems, it immediately follows that such an AI system could modify its own ethics. Not only is this appealing from a computational creativity standpoint,[7] but also it admits the potential for an agent to avoid various Asimovian paradoxes that result when an agent possesses a fixed (normative) ethics.

Additionally, the implication is that we then should allow (and even welcome) AI systems that employ as their behavioral aesthetic *any* (or any combination of) normative ethics that is valued by the meta-level-ethics-based aesthetic. Creative norms produced in this way should be valued for their novelty and value and could even possibly inform human ethics.

What would a meta-level ethics look like? While a full treatment of this question is beyond the scope of this paper, we offer a few possible starting points for such a discussion. At the base level, we could directly appeal to extant and specific ethical systems—the Golden Rule, Kant's principle of treating others as *ends*, the utilitarian principle of *maximizing the overall good*, or the categories of duty from intuitionism. Unfortunately, it is less clear what the more abstract analogs would be for candidates to be operationalized as a meta-level aesthetic. We might consider as a starting point something like a principle of consistency—a normative ethics should treat similar situations similarly.

Another possibility might be an attempt at operationlizing something like what has been called a "reflective equilibrium." Such an approach, first suggested by John Rawls, tries to find some sort of balance between the *principles* we accept and the *intuitions* about particular cases we encounter. A CC system might construct a model of common human intuition (whether about trolley problem cases or other more common cases) through some type of inductive learning. The Moral Machine[8] project at MIT is an attempt to do exactly this for the specific case of self-driving cars. Employing such a model, the system could produce a normative ethics that is responsive to this (modeled) reflective equilibrium. As an example of how such a reflective equilibrium might work, take something like the Trolley Problem. As Michael Sandel puts it:

> One principle that comes into play in the trolley story says that we should save as many lives as possible, but another says it is wrong to kill an innocent person, even for a good cause. Confronted with a situation in which saving a number of lives depends on killing an innocent person, we face a moral quandary. We must try to figure out which principle has the greater weight, or is more appropriate under the circumstances (Sandel 2009, p. 24).

Here, we must balance the utilitarian principle of *save as many lives as possible* with the deontological principle of

[6] It solves the questions, assuming, of course, some viable representation for normative ethics and some appropriate and operationalizable meta-level ethics.

[7] It has been suggested that the ability to change one's own aesthetic is critical for autonomous creativity (Jennings 2010).

[8] http://moralmachine.mit.edu

avoid harming innocent people, even for a good cause. To do so, we look for a principle that takes into account our intuitions on the subject. There is thus an interaction between our principles and intuitions that (hopefully) results in better principles. Anderson, *et al.* make a similar point in an essay, in which they write:

> Such an approach hypothesizes an ethical principle concerning relationships between [our] duties based upon intuitions about particular cases and refines this hypothesis as necessary to reflect our intuitions concerning other particular cases. As this hypothesis is refined over many cases, the principle it represents should become more aligned with intuition (Anderson, Anderson, and Armen 2005).

Considering the task of teaching ethics to (human) students provides another point of view. Elsewhere it has been argued that when we teach ethics to students, we need to focus on principles that are common to all major moral theories—since what we ought to do (for many common ethical decisions) will be answered in a similar way even by differing moral theories. For example, one of the best ways to teach ethics is to attempt to articulate

> some of the fundamental moral intuitions and principles found in almost all moral theories—for example, that all persons deserve respect and that there are minimal standards in terms of which we all expect others to treat us and which we in turn can be expected to treat others, and so on. The important thing is to articulate claims that most students should find fairly intuitive in order to strengthen their sense that there are universally valid, moral principles. The point is not that there are easy answers or absolute rules to determine every ethical decision, but rather to show students that there are moral principles that extend beyond individual preference, and across contexts, and can guide us in making such decisions (Gates, Agle, and Williams 2018).

Applying this to our meta-level CC system, if we can find common abstractions across multiple (base-level) normative ethics, and if we can formalize those abstractions we will have the basis for a reasonable approach to meta-level ethics that should produce normative ethics that will be generally accepted.

Evaluation

Supposing we could build the hybrid base-meta-level AI system for ethical behavior, how would we evaluate it? This can be addressed in multiple ways. First, from a CC point of view, we would want to know if the system is *creative*. How to establish this is still an open question, but there are several approaches to evaluation of CC systems that have been proposed. Collectively, these can examine both system product and process and include Ritchie's suggestions for formally stated empirical criteria focusing on the relative value and novelty of system output (2007); the FACE framework for qualifying different kinds of creative acts performed by a system (Colton, Charnley, and Pease 2011); the SPECS methodology which requires evaluating the system against standards that are drawn from a system specification-based characterization of creativity (Jordanous 2012); and Ventura's proposed spectrum of abstract prototype systems that can be used as landmarks by which specific CC systems can be evaluated for their relative creative ability (2016).

Second, from a behavioral point of view, we would want to know a) if the system's behaviors are *ethical* and b) if the system's behaviors are *useful*. Given that the main argument here concerns ethical behavior, the former must be the point of focus, but, given that, the latter will bear evaluation as well. Evaluating the ethics of such system behaviors is no more or less difficult than it is with extant AI systems or with humans.[9] Evaluating the utility of system behaviors is a well-understood problem and can be addressed using traditional AI evaluation methods, given a particular measure of utility.

Third, from an ethical point of view, we would want to *comprehend* the ethics of the system. Interestingly, given that the proposed system includes a meta-level for inventing normative ethics, this suggests the idea of developing a descriptive ethics for such AI systems. For obvious reasons, this is likely to be somewhat easier than doing so for human subjects, and at the same time, it is possible that the empirical study of populations of ethical AI systems could shed light on human ethics as well. For example, it is not difficult to imagine a large population of agents, all of whom possess the same meta-level ethics, admitting an empirically derived, potentially comprehensive description of that meta-level ethics. If that meta-level ethics is an operationalization of a cognitively plausible approach to ethics, one *might* be able to draw dependable conclusions about a human population operating under the meta-level ethics in question. Or, we might imagine scenarios involving multiple groups of agents, where each group possesses a different meta-level ethics, admitting the possibility of *differential* descriptive ethics that would likely be impossible with human subjects yet might yield conclusions that at least partially translate to such subjects.

Additional Considerations

There are many other interesting angles to consider here. For example, so far we have implicitly assumed that it is possible to create a domain-independent ethics. That is, given a meta-level ethics, an agent can use this as an aesthetic for creating a normative ethics that can then be applied as an aesthetic for judging candidate actions, *independent of the domain in which those actions may be applied*. The reality of *applied* ethics suggests that this assumption is likely incorrect—that rather than having a meta-level system that creates normative ethics, we should be thinking about a meta-level system that creates applied ethics. This means that the agent's environment (in a very general sense) must somehow inform either the aesthetic or the meta-aesthetic (or possibly both). Perhaps the meta-level can still produce a normative ethics and the base-level aesthetic can somehow specialize this appropriately for the domain of application. Or, perhaps the

[9] That is to say, this is likely even more difficult than addressing the question of the system's creativity.

meta-aesthetic must incorporate the domain of application, producing directly an applied ethics as its output artifact. It is, of course, possible that the same concern applies at the meta-level and that we can not even hope for a domain-independent meta-level ethics, but for now we will ignore this.

Another interesting consideration is the social aspect of ethics. Jennings makes a rather elegant argument about the social aspects of creativity and how, somewhat paradoxically, autonomous creativity *requires* significant social interaction (2010). Because his arguments center on the aesthetic judgment of the agent, they can be somewhat readily applied to our current discussion. He proposes that an agent in a social setting will not only have a model of its own aesthetic but also will have a model of its beliefs about other agents' aesthetics; it is in the dynamic updating of these models, due to social interactions, that the agent can develop true autonomous creativity; and, these social interactions are driven by psychologically plausible mechanisms such as propinquity, similarity, popularity, familiarity, mutual affinity, pride, cognitive dissonance, false inference and selective acceptance seeking. Because we are proposing ethics as aesthetic, we can follow a similar train of thought—an agent can model not only its own ethics but also (its perception of) those of all other agents. Social interaction can be a driving force behind the evolution of ethics, both at the individual and at the group level.

Yet another area for further study is the computational tenability of the proposed approaches. There is a rather simple argument for why the general problem of CC may not be computable that hinges on the decidability of the aesthetic (Ventura 2014). If the aesthetic *is* decidable, then the problem of generating candidate artifacts and filtering them with the aesthetic is computable (though efficiency could certainly still be an issue); however, if the aesthetic is *not* decidable, there is a simple reduction from the halting problem that shows that the creation of artifacts is not computable (in the theoretical computer science sense). This means that any operationalized ethics or meta-level ethics must be decidable, and given the nature of ethics, it is not clear how onerous a requirement this may be.[10]

Conclusion

We've proposed an appeal to computational creativity that addresses the problem of ethical agent behavior, which to our knowledge is a new way to look at the problem—suggesting a base-level system for which ethics is employed as an aesthetic for selecting behaviors coupled with a meta-level system for which meta-level ethics is employed as a meta-aesthetic for selecting ethics. This approach is, additionally, a new application of computational creativity, as, to date, no systems have been proposed for creating in the abstract domain of general behavior, nor, in particular, in the domain of ethics. While the current work is a position statement that asks many more questions than it answers, we believe the ethics-as-aesthetic approach to the problem of ethical agent behavior offers at least one, and possibly the only, way forward.

[10] Is it possible that recognizing an ethical action is "easy" while recognizing an unethical action is "hard"? Perhaps society itself accepts as ethical those actions that everyone deems ethical and rejects as unethical those that no one deems ethical but isn't sure about those with mixed reception. Any operationalized ethics that accurately models such a scenario will not be decidable given the existence of all three types of action.

References

Anderson, M., and Anderson, S. L., eds. 2006. *Special Issue on Machine Ethics*, volume 21(4). IEEE Intelligent Systems.

Anderson, M., and Leigh, S., eds. 2011. *Machine Ethics*. Cambridge University Press.

Anderson, M.; Anderson, S.; and Armen, C. 2005. Towards machine ethics: Implementing two action-based ethical theories. Technical report, *AAAI Fall Symposium*. 1-7.

Asimov, I. 1950. *I, Robot*. Bantam Books.

Audi, R. 2009. *Business Ethics and Ethical Business*. Oxford University Press.

Bickerman, G.; Bosley, S.; Swire, P.; and Keller, R. M. 2010. Learning to create jazz melodies using deep belief nets. In Ventura, D.; Pease, A.; Pérez y Pérez, R.; Ritchie, G.; and Veale, T., eds., *Proceedings of the International Conference on Computational Creativity*, 228–237.

Binsted, K., and Ritchie, G. 1994. A symbolic description of punning riddles and its computer implementation. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 633–638.

Bostrom, N., and Yudkowski, E. 2014. The ethics of artificial intelligence. In Frankish, K., and Ramsey, W. M., eds., *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. 316–334.

Coeckelbergh, M. 2014. Sure, artificial intelligence may end our world, but that is not the main problem. *WIRED*.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*, 21–26. IOS Press.

Colton, S.; Bundy, A.; and Walsh, T. 1999. HR: Automatic concept formation in pure mathematics. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 786–791.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., and D'Inverno, M., eds., *Computers and Creativity*. Berlin, Germany: Springer-Verlag. 3–38.

Cook, M.; Colton, S.; and Gow, J. 2016. The ANGELINA videogame design system, part I. *IEEE Transactions on Computational Intelligence and AI in Games* to appear.

European Parliament, Committee on Legal Affairs. 2017. *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*. European Commission. Retrieved January 12, 2017.

Fox, S. 2009. Evolving robots learn to lie to each other. *Popular Science*.

Gates, D.; Agle, B. R.; and Williams, R. N. 2018. Teaching business ethics: Current practice and future directions. In Heath, E.; Kaldis, B.; and Marcoux, A., eds., *The Routledge Companion to Business Ethics*. Routledge. 60–76.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Jones, T. 1991. Ethical decision making by individuals in organizations: An issue-contingent model. *The Academy of Management Review* 16(2):366–395.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Kagan, S. 1998. *Normative Ethics*. Oxford: Westview Press.

Kant, I. 1994. *Ethical Philosophy*. Indianapolis: Hackett. Trans. James W. Ellington.

Kurzweil, R. 2005. *The Singularity is Near*. Penguin Books.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2012. Adapting models of visual aesthetics for personalized content creation. *IEEE Transactions on Computational Intelligence and AI in Games* 4(3):213–228.

Lin, P.; Bekey, G.; and Abney, K. 2008. *Autonomous Military Robotics: Risk, Ethics, and Design*. US Department of Navy, Office of Naval Research.

McGee, G. 2007. A robot code of ethics. *The Scientist*.

Morris, R.; Burton, S.; Bodily, P.; and Ventura, D. 2012. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *Proceedings of the 3rd International Conference on Computational Creativity*, 119–125.

Muehlhauser, L., and Helm, L. 2012. Intelligence explosion and machine ethics. In Eden, A.; Søraker, J.; Moor, J. H.; and Steinhart, E., eds., *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Berlin: Springer.

Müller, V. C. 2016. *Risks of Artificial Intelligence*. CRC Press - Chapman & Hall.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *Journal of Creative Behavior* 47(2):106–124.

Oliveira, H. G. 2012. PoeTryMe: a versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

Özbal, G.; Pighin, D.; and Strapparava, C. 2013. BRAINSUP: Brainstorming support for creative sentence generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1446–1455.

Pachet, F., and Roy, P. 2014. Non-conformant harmonization: the real book in the style of Take 6. In *Proceedings of the 5th International Conference on Computational Creativity*, 100–107.

Pérez y Pérez, R., and Sharples, M. 2004. Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based Systems* 17(1):15–29.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39(1):217–268.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Sandel, M. 2009. *Justice: What's the Right Thing to Do?* Farrar, Straus and Giroux.

Smith, M. R.; Hintze, R. S.; and Ventura, D. 2014. Nehovah: A neologism creator nomen ipsum. In *Proceedings of the 5th International Conference on Computational Creativity*, 173–181.

Stock, O., and Strapparava, C. 2003. HAHAcronym: Humorous agents for humorous acronyms. *Humor - International Journal of Humor Research* 16(3):297–314.

Strapparava, C.; Valitutti, A.; and Stock, O. 2007. Automatizing two creative functions for advertising. In *Proceedings of 4th International Joint Workshop on Computational Creativity*, 99–105.

Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the 3rd International Conference on Computational Creativity*, 175–179.

Varshney, L.; Pinel, F.; Varshney, K.; Schorgendorfer, A.; and Chee, Y.-M. 2013. Cognition as a part of computational creativity. In *Proceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing*, 36–43.

Veale, T., and Hao, Y. 2007. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 1471–1476.

Veale, T. 2013. Less rhyme, more reason: Knowledge–based poetry generation with feeling, insight and wit. In Maher, M. L.; Veale, T.; Saunders, R.; and Bown, O., eds., *Proceedings of the Fourth International Conference on Computational Creativity*, 152–159.

Ventura, D. 2014. Can a computer be lucky? and other ridiculous questions posed by computational creativity. In *Proceedings of the Seventh Conference on Artificial General Intelligence*, 208–217. LNAI 8598.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In Pachet, F.; Cardoso, A.; Corruble, V.; and Ghedini, F., eds., *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24.

Ventura, D. 2017. How to build a cc system. In *Proceedings of the 8th International Conference on Computational Creativity*, 253–260.

Wallach, W., and Allen, C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. USA: Oxford University Press.