Minds & Machines (2010) 20:489–501
DOI 10.1007/s11023-010-9206-y

Developing Creativity: Artificial Barriers in Artificial Intelligence

Kyle E. Jennings

Received: 16 October 2009 / Accepted: 29 March 2010 / Published online: 2 October 2010

Abstract The greatest rhetorical challenge to developers of creative artificial intelligence systems is convincingly arguing that their software is more than just an extension of their own creativity. This paper suggests that "creative autonomy," which exists when a system not only evaluates creations on its own, but also changes its standards without explicit direction, is a necessary condition for making this argument. Rather than requiring that the system be hermetically sealed to avoid perceptions of human influence, developing creative autonomy is argued to be more plausible if the system is intimately embedded in a broader society of other creators and critics. Ideas are presented for constructing systems that might be able to achieve creative autonomy.

Keywords Computational creativity · Autonomy · Socially-inspired computing

The Quest for Creative Autonomy

Much of the theoretical work in creative artificial intelligence tries to specify when a system has gone beyond simply doing the bidding of its programmer. For instance, one rationale for Boden's (1991) "transformational" criterion is that since the programmer creates the initial search space with a particular view of what is possible, a system that transformed that space would be going beyond the programmer's vision. Similarly, Ritchie's (2007) "inspiring set" helps determine whether an idea produced by the system was directly involved in the system's creation or training. Finally, Colton's (2008) inclusion of imagination in his creative tripod hearkens to the autotelic exploration of ideas that is not tethered to outside forces.

K. E. Jennings (✉)
Institute of Personality and Social Research, University of California, Berkeley, CA, USA
e-mail: jennings@berkeley.edu

A computer program that relies on human judgments during the act of creation can justifiably be seen as an extension of the operator's creativity. To move a step beyond, the judgments that the human operator contributes must be encoded or learned so that the system can function independently. However, such a system would be like an apprentice who slavishly adheres to the master's creative sensibilities. While such an apprentice may be creative at some basic level, fuller creative achievement requires developing ideas about what is valuable or interesting that move beyond the master's standards. Thus, for creative artificial intelligence to progress from a capable apprentice to a creator in its own right, it must be able to both independently apply and independently change the standards it uses. This ideal will be called "creative autonomy," and represents the system's freedom to pursue a course independent of its programmer's or operator's intentions.

Once a program has started executing, any contact between the system and a human operator can lead to justifiable questions about the system's alleged independence. One reaction might be to hermetically seal the system from the outside world. However, human creativity takes place within a rich web of social interactions (Csikszentmihalyi 1988; Sawyer 2007), a fact that does not call into question any one person's potential to be truly creative. In fact, it is in making sense of and responding to interactions with other creators that we arrive at a style that is unique to us, yet not so unusual that others will not take it seriously. Creative autonomy will likewise be argued to emerge out of the interactions with multiple critics and creators, not from solitary confinement.

Definition Creativity is a social construction, and thus cannot be reduced to unassailable formal properties. Still, it is useful to suggest conditions under which a system may more readily be seen as creative, so that they may be investigated theoretically and empirically. Toward that end, a system will be said to have creative autonomy if it meets the following three criteria:

Autonomous Evaluation—the system can evaluate its liking of a creation without seeking opinions from an outside source
Autonomous Change—the system initiates and guides changes to its standards without being explicitly directed when and how to do so
Non-Randomness—the system's evaluations and standard changes are not purely random

Autonomous evaluation requires that the system be able to issue opinions without consulting an outside human or machine intelligence. However, the system is free to ask for or observe others' opinions at other times, and to store this information. Autonomous evaluation could easily be achieved by using preprogrammed standards or by learning another source's standards, both of which could be used to bootstrap a system. After this, however, *autonomous change* requires that the system be able to independently change its standards. Though external events may prompt and guide changes, the system cannot exclusively rely on another source to tell it when to change standards, or when its new standards are acceptable, nor can it simply make some fixed transformation to another source's standards.

An easy way to satisfy both criteria would be to issue random decisions, or to make random criteria changes at random times. The *non-randomness* requirement is meant to prevent this. Many algorithms incorporate randomness, so not all randomness is precluded. For instance, the system could resolve conflicts between standards randomly, or it could test random perturbations to its standards. Aside from special cases like these, however, it cannot simply make random decisions. Of course, this does not guarantee predictable outcomes.

Relationship to Other Concepts

Autonomy and Agency The term "autonomy" is often used in connection with software agents, which Wooldridge and Jennings (1995) define as having autonomy and social ability, and being able to act both reactively and proactively. There are two broad kinds of autonomy that are discussed in relation to agents: executive autonomy, and goal autonomy (Castelfranchi 1995). A system with executive autonomy can choose how to achieve an externally-provided goal, while a system with goal autonomy can also decide whether to accept that goal. Similarly, a system capable of autonomous evaluation can apply externally-provided standards, while a system capable of autonomous change can also decide how to change these standards.

Though autonomous evaluation and change are analogous to executive and goal autonomy, they are not the same. In particular, since creative autonomy refers only to a system's evaluation standards, it can apply to a critic (a system that can evaluate creations, but that can't itself create). Little is gained by speaking of this critic's goals, just as little is gained by speaking of a scale's or a ruler's goals (cf. Shoham 1993, p. 53). For systems that can create, it becomes meaningful to ascribe goals (e.g., to create something it likes), but goal autonomy and autonomous change are still separate. For instance, a creative system that produces music according to tastes it arrived at independently is no less impressive if it cannot one day decide to deny its operator's request to compose a jovial melody.

Given this discussion, it is clear that a system with creative autonomy won't necessarily have the autonomy discussed by Wooldridge and Jennings. This makes creative autonomy an orthogonal issue to agency. As a consequence, autonomously creative systems can potentially retain the reactive nature of ordinary software. Though it will be argued that achieving creative autonomy requires contact with multiple outside intelligences, such contact is not part of the definition of creative autonomy.

Creativity versus Creative Autonomy Creativity is evaluated both via *what* is produced, and *how* it was produced (Kasof 1995). Creative autonomy may prove to be a necessary (though perhaps not sufficient) condition for the "how", but does not guarantee anything about the "what". A system with creative autonomy can be seen as having the potential to do creative things, but it is a separate question whether it ever will. Even if the system does *do* something creative, it could turn out that other criteria will apply to the "how" (possibly including those in the definition of "agent") before the system itself can be considered to *be* creative. Still, creative

autonomy is a useful concept and an important first step toward achieving machine creativity.

Sources of Change

In highly formalized domains, it may be possible to hand code evaluation criteria. In other cases, a system may do better to learn by example. In either case, criteria changes must occur without explicit instruction from the programmer or expert. In principle, these changes could occur endogenously (without input from outside the system) or exogenously (based on input from outside the system), with separate implications for creative autonomy.

Suppose that the system's changes are entirely endogenous. For instance, it might promote rules that lead to more efficient searches, or eliminate rules that lead to contradictory conclusions. Though this system's criteria would change in unpredictable ways, they are still probabilistically "preordained," in that the probability distribution of outcomes is set before the program starts. What's more, since a *field's* evaluation criteria change in light of new contributions (Sawyer 1999), the system is in danger of growing obsolete, or of straying too far from other creators to be taken seriously.[1]

As Castelfranchi (1995) observes, isolation is not a true form of autonomy. A more palatable and realistic option is for the system to consider exogenous information as it changes its criteria. They key question is how a system can maintain this external connection without being slave to the information it receives.

The remainder of this paper considers how an externally attuned creative system might achieve autonomous change without merely tracking changes in one or more external authorities unconditionally. The sketched solution involves a set of considerations that are orthogonal to evaluation criteria, but that have meaningful connections to how human creators develop and change. The paper closes with a discussion of implications and open questions.

Learning to Evaluate

This section describes how a creator could learn evaluation standards via its interactions with others. Since both creators and non-creators have opinions, these "others" will be called "critics" rather than "creators". (All creators are critics, but not all critics are creators.) The notation used below is based on Wiggins' model of creative search (Wiggins 2006). However, the only thing of interest here is how a creator learns to judge the quality of finished products. How these standards affect the search process, as well as how they apply to intermediate products, are left as interesting questions for future work. This model assumes that there is at least one critic, though its more interesting features do not apply unless there are more.

[1] If the programmer considers the experiment to have failed and manually alters the system's code or knowledge, the new version should be considered an extension of the old, in which case autonomous change was not achieved.

Though the processes described here are inspired by human creativity, they could be implemented in a society of solely AI creators, or in a mixed human-machine society.

Subjectivity Assume that a creation's value is at least in part socially constructed, and that different critics have different standards. Wiggins uses $\mathcal{E}$ for the knowledge representing these standards, which can be subscripted to indicate whose standards are in question, e.g., $\mathcal{E}_i$.[2]

In addition to knowing that people have different opinions, we can often estimate a typical or specific person's opinion (Csikszentmihalyi and Sawyer 1995; Fourquet-Courbet et al. 2008). Thus, the knowledge in $\mathcal{E}_i$ can be segmented by whose evaluations the knowledge is about. For this we will use the subscript ij, where i is the perceiver and j is whose opinion is perceived. A dot ("·") will represent the typical critic. Thus,

$$\mathcal{E}_i = \langle \mathcal{E}_{i\cdot}, \mathcal{E}_{i1}, \ldots, \mathcal{E}_{ii}, \ldots, \mathcal{E}_{iN} \rangle$$

where N is the number of critics in the society. Knowing other critics' preferences lets a creator target an audience, and so it is important for the information to be correct. Therefore, assume that creators continuously correct inaccuracies.

For sake of argument, assume that creators represent knowledge at the most general level possible and avoid duplicating knowledge. This means that something that applies to most creators would be stored in $\mathcal{E}_{i\cdot}$, and creator j's deviations from this would be stored in $\mathcal{E}_{ij}$. If creator i knows nothing specific about creator j's standards, then $\mathcal{E}_{ij} = \emptyset$. We will assume the most difficult case in which creators start with no standards of their own, i.e., $\mathcal{E}_{ii} = \emptyset$.

Making Evaluations Creator i's evaluation of creation c from critic j's perspective is denoted by $E_{ij}(c)$. Though the notation is analogous, E_{ij} is not simply the application of $\mathcal{E}_{ij}$. This is because the applicability of $\mathcal{E}_{ij}$ depends on c. For instance, if the discipline is furniture design, creator i might know a great deal about how j evaluates chairs, but nothing about how j evaluates tables. If c is a table, it would makes more sense for i to rely on $\mathcal{E}_{i\cdot}$ than $\mathcal{E}_{ij}$ when estimating j's evaluation.

Wiggins writes $[\![\mathcal{X}]\!]$ to mean the translation of the knowledge in $\mathcal{X}$ (where $\mathcal{X}$ is $\mathcal{R}$ or $\mathcal{E}$) to a function from creations to real numbers in [0, 1]. We will extend this to map to [0, 1] × [0, 1], for the result and the confidence in that result. Additionally, we need a function that can aggregate different evaluations and confidence levels into a single answer. Heuristics such as assuming that people from similar backgrounds have similar opinions could compensate for missing information. Such details don't matter here, and so we'll simply say that each creator has background social knowledge, $\mathcal{S}_i$, and a function F_i that uses this knowledge to consolidate the other information. Thus, for $i \neq j$ we have:

[2] Wiggins uses $\mathcal{L}$ to denote the language that this knowledge is expressed in. Such details do not matter in the present treatment, and so no assumptions will be made about representation.

$$E_{ij}(c) = F_i(\mathcal{S}_i, j, [\![\mathcal{E}_{i\cdot}]\!](c), [\![\mathcal{E}_{i1}]\!](c), \ldots, [\![\mathcal{E}_{iN}]\!](c))$$

In the extreme case, a creator's own opinion would only depend on knowledge in $\mathcal{E}_{ii}$. However, by hypothesis $\mathcal{E}_{ii}$ is initially empty, meaning that the creator must construct its opinion from what it knows about others' opinions. Though we could make the system issue the most representative opinion, it will prove more interesting if the system prefers to emulate some critics more than others, based on its affinity for each critic. These affinity levels are stored in $\mathcal{A}_i$, and are discussed in the next section. We can now define an analogous function to F_i:

$$E_{ii}(c) = F'_i(\mathcal{S}_i, \mathcal{A}_i, [\![\mathcal{E}_{i\cdot}]\!](c), [\![\mathcal{E}_{i1}]\!](c), \ldots, [\![\mathcal{E}_{iN}]\!](c))$$

Note that E_{ii} would be just one component of the creator's objective function during search (cf. Jennings 2008), but is the only function creator i uses to evaluate its own and others' finished products.

Communication Creator i learns to make autonomous evaluations via interactions with other critics. Suppose that a creator i has made a creation c, which is observed by a critic $j \neq i$. There are three broad classes of information that can be communicated.

Evaluation—A simple "like/dislike" judgment. Critic j communicates $E_{jj}(c)$, and then creator i adjusts its knowledge until $E_{ij}(c) \approx E_{jj}(c)$.
Correction—Critic j creates c', a modification of c that it likes better. Creator i updates its knowledge so that $E_{ij}(c') > E_{ij}(c)$, and tries to determine what changes between c and c' increased j's liking.
Criticism—Justifications for an evaluation or correction, e.g., what is pleasing or what criteria were used. Critic j communicates knowledge in or derived from $\mathcal{E}_j$ to creator i, which attempts to integrate this knowledge into $\mathcal{E}_i$. If i cannot make $E_{ij}(c) \approx E_{jj}(c)$, then i might ask j for clarification.

In each case, creator i adjusts $\mathcal{E}_i$ in order to reproduce j's evaluation. Because knowledge is represented at the most general level and duplication is avoided, this adjustment should always result in change to $\mathcal{E}_{i\cdot}$ or to $\mathcal{E}_{ij}$. These processes cannot by themselves make $\mathcal{E}_{ii}$ non-empty.

In creative AI systems that only allow interaction with one critic (e.g., the programmer), all of the system's knowledge can be represented in $\mathcal{E}_{i\cdot}$, meaning that $E_{ii}(c) = E_{ij}(c) = E_{i\cdot}(c) = [\![\mathcal{E}_{i\cdot}]\!](c)$, i.e., the system parrots back its understanding of the critic's standards. The situation improves somewhat with multiple critics since the system forms $E_{ii}(c)$ from many different sets of standards in ways dependent on $\mathcal{S}_i$ and $\mathcal{A}_i$. However, it still only offers direct translations of other critics' standards. What's more, in both cases, the system's need to represent other creators' standards faithfully implies that its standards only change in reaction to and in proportion to changes in other critics' standards. Hence, though these processes support autonomous evaluation and are non-random, they are not enough for creative autonomy. The next section suggests some extensions that would add the missing component, autonomous and non-random change.

Changing Standards

If the system faithfully updates its knowledge of others' standards, autonomous change will not occur until there is knowledge in $\mathcal{E}_{ii}$, which is the only knowledge that can evolve independently of other creators' standards. Since all of the system's knowledge comes from other critics and is stored at the most general level, there is as yet no reason for knowledge to enter $\mathcal{E}_{ii}$. Inspired by human psychological processes that would be relatively simple to implement, this section suggests some reasons that $\mathcal{E}_{ii}$ might be initially populated and subsequently changed.

Additional Behaviors

As described so far, the system combines others' preferences according to how applicable the preferences are and how much the system "likes" each critic. This section first describes how "liking" could initially be configured and then changed. Next, the system is given reasons to doubt its own evaluations, which thus far it has never had cause to do. This doubt will later be argued to lead to including knowledge in $\mathcal{E}_{ii}$.

Affinity Any number of rules could be used to set the initial affinities in $\mathcal{A}_i$, all of which have a basis in human psychology:

Propinquity—Our friendships (Nahemow and Lawton 1975) and collaborations (Kraut et al. 1988) are largely determined by physical proximity. Analagously, the system could initially be set to prefer creators who are nearby in some topology.
Similarity—We subconsciously favor people with similar backgrounds. In a society of artificial creators with varied parameterizations, similarly parameterized creators might initially prefer each other.
Popularity—When we cannot make sense of a speaker's message, we decide whether to believe her based on cues about her prestige (Cialdini 2007), e.g., age (time in the society) or popularity (received affinity).

Some affinity changes would be independent of the system's evaluations:

Familiarity—Absent other discernable differences, we tend to prefer people and things we have seen before (Zajonc 1968). Frequent interactions could increase liking.
Mutual Affinity—We are more apt to like someone if they first show that they like us (Mettee and Aronson 1974). The system could increase its affinity for critics that evaluate the system's creations positively.

Finally, affinity could adjust in response to the creator evaluating a critic's work, or by how closely the creator and critic agree on evaluations of a third creator's work.

Self-Confidence and Pride Unsure about the quality of their work, novices are particularly sensitive to praise and criticism. The sting of failure can be offset by the memory of success (Heine et al 2006), helping the novice maintain self-confidence.

A person whose self-confidence is too low for too long is likely to give up and pursue other interests.

A creative system could be written with a need to maintain a certain level of self-confidence. While this need might seem like an unnecessary liability, it also provides a rationale for many kinds of actions. In particular, a system might maintain memories of proud accomplishments as a way to preserve self-confidence in the face of failure.

Suppose that a system has a memory of past successes, $\mathcal{M}_i$, and their last average evaluation, $M_i = \sum_{c \in \mathcal{M}_i} E_{ii}(c)/|\mathcal{M}_i|$. Only highly salient creations would be stored (ones that elicited "pride"), such as creations that got unexpectedly high evaluations (relative to recent creations, other creators' creations, or the critic's typical evaluation), particularly from a critic the creator likes. As with a person who concludes that all of her prior work was worthless, there could be negative repercussions for the system if the value of M_i suddenly dropped, particularly if other sources of self-confidence (e.g., recent external approval) were also lacking. As discussed next, avoiding this drop could lead the system to develop its own standards.

Bootstrapping and Changing $\mathcal{E}_{ii}$

This section introduces three processes that could introduce and change knowledge in $\mathcal{E}_{ii}$. As before, each is inspired by human behavior. The processes are sketched here, and discussed relative to creative autonomy in the next section.

Cognitive Dissonance Consider a human novice whose evaluations mirror an influential mentor's, and whose self-confidence rests on memories of his past successes. Suppose that one particular creation, which was highly rated by his mentor, is a large source of pride. He only understands why that work was good in terms of how he understands his mentor's preferences, which he trusts since he respects that mentor. Now suppose that the mentor strongly criticized a highly similar creation, throwing into doubt his understanding of the mentor's standards. Or, perhaps an unrelated event would make him lose respect for the mentor, leading him to discount the mentor's opinion. In either case, he could no longer justify such a high evaluation for his prized creation, leading to a dilemma: believe the reduced evaluation, and hence that he's not as good as he thought; or, doubt the new evaluation, and continue to believe he and his work are great. This "cognitive dissonance" (Festinger 1957) is distressing enough have physiological correlates (Croyle and Cooper 1983), and can lead us to alter the truth or our memories in order to allay it.

A system programmed to "feel proud" could face a similar situation. When a creation enters $\mathcal{M}_i$, the creator agrees with the critic's evaluation, that is, $E_{ii}(c) \approx E_{jj}(c)$, which, if $\mathcal{E}_{ii} = \emptyset$, was arrived at via other critics' preferences. When knowledge of these preferences or their weighting changes, some evaluations in $\mathcal{M}_i$ could drop, as would M_i. By construction, the system cannot tolerate too large of a drop. Since the system must also accurately represent others' preferences, it cannot

simply refuse to change that knowledge. To resolve this conflict, it could add information to $\mathcal{E}_{ii}$ that keeps M_i from dropping too much.

False Inferences About Preferences Criticism includes the reasons behind an overall evaluation. However, the reasons we offer do not always reflect how we make our decisions. For instance, people will say why they preferred one of many products, all of which are actually identical (Wilson and Nisbett 1978). Similarly, we invent reasons that sound good if our real reasons aren't socially acceptable. For instance, though our evaluations of one aspect of a person pollute our evaluations of other aspects (the "halo effect"; Thorndike 1920), we often don't know or admit this. Instead, we offer reasons that are demonstrably unrelated to our actual evaluation process (Nisbett and Wilson 1977).

A creator whose standards are based solely on other people's standards is unlikely to say that she likes something "because he likes it too." Instead, she will search for distinguishing features of the item to form a plausible-sounding explanation. Even if incomplete or incorrect, this utterance becomes part of how she understands her preferences, and might even impact future evaluations. Additionally, the very fact that she has have chosen something may motivate her to raise the chosen alternative's perceived superiority over similar alternatives (Brehm 1956).

Suppose that a creative AI had a language for communicating criticism. Supposing that $\mathcal{E}_i$ consists of exemplars, neural networks, and other irregular representations, there is a large chance that the language could not express complete and correct information. If the rules it extrapolates are put into $\mathcal{E}_{ii}$, two things happen. First, the inaccuracy of the rules will lead to evaluations that no longer directly follow $\mathcal{E}_i \setminus \mathcal{E}_{ii}$. Second, the creator's standards will lag behind changes in $\mathcal{E}_i \setminus \mathcal{E}_{ii}$, since those will not be reflected in $\mathcal{E}_{ii}$. Thus, the system will begin to develop divergent standards, albeit clumsily.

Selective Acceptance Seeking Even someone with a completely independent sense of what he likes might want a style somewhat similar to people he admires. If one such peer's preferences shifted, he might adjust his own preferences in that direction. However, there would likely be several other peers whom he wishes to be somewhat near to, leading to experimentation until an equilibrium is reached. Other strategies for maintaining distinctiveness while still feeling part of a group can also be imagined (see, e.g., Hornsey and Jetten 2004).

Once a creative AI relies substantially on $\mathcal{E}_{ii}$, changes in other critics' preferences will have a smaller impact on E_{ii}. However, the system might try to keep an acceptably low discrepancy between $E_{ii}(c)$ and $E_{ij}(c)$, where j is a critic whom i has a high affinity for. Indeed, this tuning might be what enables the system to deviate from others' standards but stay recognizably within the same domain or genre.

Autonomy Revisited

Creative autonomy requires autonomous evaluation, autonomous change, and non-randomness. A system such as was described above could certainly be capable of autonomous and non-random evaluation. Furthermore, none of the schemes

described above makes random changes at random times (though admittedly the described changes would be more accidental than deliberate). Therefore, it just remains to be considered whether the system's changes would be autonomous. Castelfranchi (1995) says that a social agent has autonomy if it isn't bound by other agents' goals for it. Similarly, the test for whether a system has creative autonomy is whether an external agent could somehow cajole the system into adopting criteria that the external agent chose for it.

In all three schemes described above, change happens in response to external events. For cognitive dissonance and acceptance seeking, these events would be changes in others' standards ($\mathcal{E}_i \setminus \mathcal{E}_{ii}$), or in affinities ($\mathcal{A}_i$), possibly only after the effect of several separate changes had accumulated. For false inferences, the event would be the request for a critique. Unless a critic could manipulate when the creator starts and stops changing its standards, these change processes could be autonomous.

With only a single critic (such as the system's creator), such manipulation is possible. Acceptance seeking would simply entail following the lone critic's changes in standards within a margin of error. The critic could take advantage of false inferences by requesting criticisms as a way to perturb the creator's standards, only stopping when an acceptable result was reached. The critic could also give extreme and inconsistent ratings to trigger changes via cognitive dissonance.

The situation changes with multiple critics. The complex web of relations between $\mathcal{A}_i$, $\mathcal{E}_i$, $\mathcal{M}_i$, and M_i would make it hard to predict whether an external change would trigger adjustments to $\mathcal{E}_{ii}$. Multiple simultaneous changes might cancel out, or several small changes across time could accumulate until one change unleashes a string of compensatory adjustments. These features would make the system less clearly responsive to any single critic, and in particular much more difficult for any single critic to manipulate.

Autonomy also precludes making fixed transformations to others' standards. When knowledge is first put into $\mathcal{E}_{ii}$, it is derived with error from others' standards, such as the stereotyped rules delivering criticism would produce, or the extrema that cognitive dissonance would enshrine. One could argue that these things would only result in time-lagged caricatures of other critics' preferences. The counter-argument is that in a society where every creator pays attention to several other creators, such distortions would serve as attractors, leading to unpredictable clusters of similar styles, and unpredictable shifts in those styles. These emergent dynamics would make it impossible to say which creator was leading the change, meaning at least that no one creator was more autonomous than another.

The foregoing claim requires more theoretical and empirical work. However, in a single-critic system, it is a fair guess that $\mathcal{E}_{ii}$ would be more of a fun-house mirror reflection of the critic than anything that could be considered autonomously arrived at. Given that it would also be impossible to rule out that the critic had manipulated the system's dynamics until such time as the system arrived at standards that the critic liked, it seems fair to say that single-critic systems, at least as sketched here, would not achieve creative autonomy. The answer for multiple-critic systems will have to wait.

Conclusions

This paper introduced the concept of creative autonomy, which requires that a system be able to evaluate its creations without consulting others, that it be able to adjust how it makes these evaluations without being explicitly told when or how to do so, and that these processes not be purely random. Extending work by Wiggins (2006), a notation was developed to denote evaluations drawn from the integration of knowledge about several different critics' standards. Importantly, the system has different affinities for each critic, which impact how it integrates their opinions to form its own opinion. Initially the system has no independently-held preferences, but this can change when the system attempts to justify its evaluations, or if the system must maintain high evaluations for some of its past work in the face of other critics' changing standards.

Such a system was argued to be capable of autonomous, non-random evaluation, and the change processes sketched are non-random. In a single critic society (i.e., one with the programmer and the software), it is unlikely that the system's standards would be more than distorted agglomerations of the critic's standards. What's more, the ease with which the critic could manipulate the system would make it hard to argue that the creator was changing autonomously. In a multiple-critic society, complex interactions might make any one creator impervious to manipulation, and emergent dynamics of the system could lead to clusters of creative styles. However, these ideas await empirical demonstration. Additionally, the philosophical question of whether changing absent direct manipulation is the same as autonomy must be answered.

The description of creative autonomy offered here captures only a small part of why humans can be considered creative. A system whose creations had a style that was not easily traced to a few influences, yet was still recognizably in the same domain, would be a major accomplishment. However, as just mentioned, freedom from manipulation is not the same as acting purposefully. The system described here only makes changes in (possibly indirect) reaction to others' changes. Human creators, in contrast, proactively change their standards. It is conceivable that this system could make proactive changes by looking for patterns in how others' standards change with time and in relation to each other, which would be proactive, though perhaps not purposeful enough to be considered creative.

Though this work does not directly address the distinction between exploratory and transformational creativity, it could lead to interesting insights in that area. In particular, transforming a search space can be seen as searching over search spaces. In addition to making transformational creativity a form of exploratory creativity, there remains the question of what objective function is used to guide the search over transformations. Repeating this question over recursive searches of transform spaces of transform spaces of transform spaces, etc., one must ask what the base case is, i.e., what is the ultimate objective function? The perspective suggested here (and doubtless elsewhere, too) is that the ultimate objective function emerges out of the interactions between creators. It thus becomes essential for any system to be able to interact fully with its creative milieu if it is to be truly creative.

Creative artificial intelligence must always fight the impression that it is simply a fancy tool for expressing the programmer's creativity. This burden can lead to a desire to isolate the system from outside influences as much as possible. However, as argued here, autonomy might require more, not less, interaction, though this interaction must extend beyond the programmer. Though the hypotheses presented here are painted with a broad brush and await verification, this work does suggest that creative AI must be viewed in a broader context than it traditionally has. Developing creative AI might still amount to solving an information processing problem, but a good part of this information comes from the social world. Of course, this is true of our own creative processes, as well.

Acknowledgments The author is grateful to Geraint Wiggins and several anonymous reviewers, whose feedback greatly improved this work, and to the organizers of and participants in the 2008 AAAI Spring Symposium on Creative Intelligent Systems and the 2009 International Joint Workshop on Computational Creativity, in discussion with whom these ideas were hatched and nurtured.

References

Boden, M. A. (1991). The creative mind: Myths and mechanisms. New York: Basic Books.

Brehm, J. W. (1956). Post-decision changes in desirability of alternatives. *Journal of Abnormal and Social Psychology, 52*, 384–389.

Castelfranchi, C. (1995). Guarantees for autonomy in cognitive agent architecture. In M. Wooldridge & N. R. Jennings (Eds.), *Intelligent agents: Theories, architectures, and languages, vol LNAI volume 890* (pp. 56–70). Heidelberg, Germany: Springer.

Cialdini, R. B. (2007). Influence: The psychology of persuasion. New York: Collins.

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *Creative intelligent systems: Papers from the AAAI spring symposium, AAAI Press, Stanford, CA* (pp. 14–20).

Croyle, R. T., & Cooper, J. (1983). Dissonance arousal: Physiological evidence. *Journal of Personality and Social Psychology, 45*(4), 782–791.

Csikszentmihalyi, M. (1988). Society, culture, and person: A systems view of creativity. In R. Sternberg (Ed.), The nature of creativity (pp. 325–339). Cambridge: Cambridge University Press.

Csikszentmihalyi, M., & Sawyer, K. (1995) Creative insight: The social dimension of a solitary moment. In R. J. Sternberg & J. E. Davidson (Eds.), The nature of insight (pp. 329–364). Cambridge, MA: MIT Press.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford, CA: Stanford University Press.

Fourquet-Courbet, M. P., Courbet, D., & Vanhuele, M. (2008). Creativity and e-advertising: A qualitative study of art directors' creative processes. *Empirical Studies of the Arts, 26*(1), 5–13.

Heine S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Personality and Social Psychology Review, 10*(2), 88–110.

Hornsey, M. J., & Jetten, J. (2004). The individual within the group: Balancing the need to belong with the need to be different. *Personality and Social Psychology Review, 8*(3), 248–264.

Jennings, K. E. (2008). Adjusting the novelty thermostat: Courting creative success through judicious randomness. In *Proceedings of the AAAI spring symposium on creative intelligent systems, AAAI Press, Stanford, CA*.

Kasof, J. (1995). Explaining creativity: The attributional perspective. *Creativity Research Journal 8*(4), 311–366.

Kraut, R., Egido, C., & Galegher, J. (1988). Patterns of contact and communication in scientific research collaboration. In *CSCW '88: Proceedings of the 1988 ACM conference on computer-supported cooperative work, ACM, New York* (pp. 1–12).

Mettee, D. R., & Aronson, E. (1974). Affective reactions to appraisal from others. In T. L. Huston (Ed.), Foundations of interpersonal attraction (pp. 235–83). New York: Academic Press.

Nahemow, L., & Lawton, M. (1975). Similarity and propinquity in friendship formation. *Journal of Personality and Social Psychology, 32*, 205–213.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84*(3), 231–259.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines, 17*(1), 67–99.

Sawyer, R. K. (1999). The emergence of creativity. *Philosophical Psychology, 12*(4), 447–470.

Sawyer, R. K. (2007). Group genius: The creative power of collaboration. New York, NY: Basic Books

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence, 60*(1), 51–92.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Wiggins, G. A. (2006). A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems, 19*, 449–458.

Wilson, T., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology, 41*(2),118–131.

Wooldridge, M. J., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*, 115–152.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2), 1–27.