Transforming Kantian Aesthetic Principles into Qualitative Hermeneutics for Contemplative AGI Agents

Jeremy O. Turner(✉) and Steve DiPaola(✉)

Simon Fraser University, Vancouver, BC V6B 5K3, Canada
{jot,sdipaola}@sfu.ca

Abstract. This paper introduces an interdisciplinary qualitative hermeneutic approach to the engineering and computer science methodological paradigm(s) for assessing a contemplative-agent's cognitive capabilities at a level corresponding to Artificial General Intelligence (AGI). This paper has utilized cognitive intensity levels from Kantian aesthetic philosophy to qualitatively re-address Russell and Norvig's canonical agent-categories as they scale upward towards AGI cognition levels. These Kantian levels allow the AGI-agent designer to consider the cognitive interplay between computational representations of the imagination and reason as they relate to motivationally-nuanced teleological notions of self-interest versus disinterestedness. While the AGI level is the thematic focus, lower intensity levels are also introduced in order to set the appropriate cognitive benchmarks for higher levels corresponding to truly contemplative AGI-agents. This paper first contextualizes Kant's analytical framework before discussing the appropriately corresponding agent-categories. This paper concludes with a brief discussion of the particular methodological and hermeneutic contribution of Kant's aesthetic philosophical framework to the AGI domain.

Keywords: AGI · Contemplation · Aesthetic philosophy · Agents

1 Introduction

Artificial General Intelligence (AGI) promises that one day, an artificially created agent will be sufficiently intelligent to be able to theoretically become competent in any computer-tractable expert domain. Interestingly, hermeneutic analysis of how such minds might become capable of interdisciplinary competency are primarily being handled methodologically by mechanistically-focused computer science and cognitive engineering disciplines. As cognitive science is an umbrella mega-discipline that includes AGI, there remains a contribution within the AGI community to try to make interdisciplinary correlations and unify hermeneutic understandings of the intelligent agent categories and their cognitive affordances with other cognitive science related disciplines. Within the related cognitive science of philosophy is the sub-discipline of aesthetic philosophy. At first glance, considering principles from a humanities-driven sub-discipline such as aesthetic philosophy would seem thematically orthogonal to the

M. Iklé et al. (Eds.): AGI 2018, LNAI 10999, pp. 238–247, 2018.
https://doi.org/10.1007/978-3-319-97676-1_23

mechanistic and technological cognitive categorization of artificially intelligent agents. However, this paper shows examples of how interdisciplinary hermeneutic correlations can indeed be made between Kantian aesthetic philosophy [1] and canonical intelligent-agent categories [2]. We proceed by first contextualizing and explaining Kant's analytical framework before discussing the appropriately corresponding agent-categories. This hermeneutic analysis concludes with a brief reflection on the particular methodological and hermeneutic contribution of Kant's aesthetic philosophical framework to the AGI disciplinary domain.

1.1 The Kantian Approach to Cognitive Intensification

Intelligent agent categories are scaled upward to indicate an agent's gradually acquired capabilities to deeply contemplate a world-model from which derived goals and values of those goals (i.e. utilities) can be learned from either through, reason, imagination, or empirical experience regardless of semantics originating from a particular task-domain. A more generally intelligent artificial agent should occasionally consider solutions to problems that reside outside of its immediate self-interest. Kant's aesthetic intensity levels are categorized according to a motivational spectrum that ranges from an agent acting in its own self-interest to engaging in disinterested contemplation for its own sake. Contemplation can be intensified in a similar way that an agent's intellectual aptitude can increase. Kant's aesthetic intensities are based on an agent being able to cognitively and rationally parse an artificially represented imagination (including having to reason about an external world-model). In this sense, Kantian aesthetic intensities are cognitive intensities involving a dynamic interplay with computational representations of imagination and reason. An agent would need to function at increasingly higher intelligence-categories to be able to mechanistically parse each intensity relating to contemplating this cognitive acceleration of imagination and reasonable estimation without functional overload and/or paralysis. The interpretation of aesthetic principles for AGI can be properly understood once the reader has internalized Kant's philosophical definitions of the imagination and reason. The cognitive interplay between philosophical faculties of imagination and reason in particular, form the basis for understanding the most contemplatively involved aesthetic intensity which can only really be properly comprehended by an AGI-agent – the transcendent sublime.

Kantian and Computational Representations of the "Imagination". The Kantian interpretation of the imagination is sometimes conflated with "intellectual intuition" [4, p. 186]. Kant's considers the imagination to be a highly qualitative and semantically generative cognitive process. The contents or inspiration of the imagination can involve: fictional, functional, or fantasy elements and can be derived from sensory empirical phenomena or from the platonic realm of pure ideas (i.e. noumena). Unlike reason, the imagination involves comprehending (understanding) or reflecting on either perceptual phenomena or ideal noumena rather than trying to practically apprehend (calculate or numerically judge) a phenomenon's or noumenon's precise properties. To imaginatively comprehend something that has entered our internal or external perception, Kant believes that we make a subjective determination of phenomenon's or noumenon's

ultimate nature and teleological purpose [4, p. 178]. Reflection is a key contemplative meta-faculty when considering the imagination because "[…] *the apprehension of forms by the imagination could never occur if reflective judgment did not compare them"* [1, pp. 29–30].

In the AI discipline, the "imagination" is defined as *"the* [contemplative] *manipulation of information* [mental imagery] *that is not directly available to an agent's sensors"* [5, p. 743]. Artificial imagination as currently implemented usually pertains to an agent's capacity to simulate a task from its problem-space into its semantic and episodic memory banks for future retrieval [5]. The current goal for handling, storing and predicting these imagined states is through the eventual implementation of a prospective memory [6]. A computational representation of imagination can include: a predicted/prospective generative percept-stream (future states, operators, probability distributions, actions etc.) and/or simulated input-output procedure that has some portion existing outside of or beyond (or in parallel to) the agent's sensory apparatus; or a form of creative-reinterpretation (or misinterpretation) of pre-existing sensory data. Imagination can be a semantically interpreted percept-stream (various parallel ontological descriptions of: states, operators, actions etc.) that has some portion existing outside of or beyond (or in parallel to) the agent's sensory apparatus. The semantic interpretation of this percept-stream might not directly address the raw data or hypotheses inferred by the agent's sensory apparatus. An agent's imagination might contain purely generative (imagined) information content for rationalization or prediction without any empirical grounding in the sensed data. Alternately, an agent might use its imagination to provide its own idiosyncratic semantic interpretation of the received empirical sense data. An agent can also use an artificially created imagination faculty to modulate its hard-coded rationality. For example, an agent can dynamically modify its ontology and its list of competency questions when expecting or predicting new percepts from its situated environment. Such dynamic self-modifications can rapidly occur even without the encouragement provided by immediate access to explicit empirical evidence. To address the imaginative context; these questions must address and/or approximate task-domain knowledge that is hypothetical, simulated (i.e. approximate symbolic representations being rehearsed in a computational imagination), or speculative (even fictional). In order to utilize the computational equivalent of an imagination, an agent's set of competency questions and expected answers should be abstracted from its observable utility in a particular environment for internal deliberation. For example, an agent might use a terminal pointer, symbol or chunking mechanism to compress recently acquired episodic data into an imagined rule procedure that could be used in a future interaction scenario. This "imagined" data can be cross-referenced with its knowledge-base (KB) and ontology [e.g. 18] in order to determine how reasonable the imagined semantic content might be. The competency questions themselves might also use the "imagination" for inspiration and be formulated without any grounded basis in pre-observed empirical data. Over time, the agent will compare imagined or simulated contemplation scenarios with those that it empirically perceives from its sensory environment. Regardless of how they are originally formulated, an agent's competency questions will be eventually formulated such that an agent can learn to semantically reify empirically sensed: subjects, predicates, intrinsic rewards and/or objects. The process of reification itself is

understood to mean the idea of transforming an imaginary concept into something "real" and therefore, something computable and reasonable. Such reifications would be continuously updated within the agent's Ontology and KB.

Kantian and Computational Representations of "Reasonable Estimation". Kant has dedicated two books towards discussing the earthly and transcendent intricacies of the faculty of reason as it relates to sensation and cognition [3, p. 1788]. The Kantian definition of reason here has been restricted to narrow the focus on an understanding of reason in terms of artificial intelligence. Kant basically felt that reason can be subdivided into two categories, "pure" (i.e. speculative) logical reason based on a priori knowledge [3, p. 15] and practical (i.e. action-centric) reason based on direct empirical experience [3, p. 267]. Logical reason could seem to mathematically estimate ideas independent of empirical experience. Kant, however, still ensured that logical reason could also be contingent on grounded representational associations with concepts and analogies drawn from empirical experience (i.e. sensation) [3, p. 132; 7, p. 25]. Anything contemplated outside of these grounded representational associations would likely be beyond the realm of pure and practical realms of reason but still well within the visually inspired realm of the imagination – which harmonizes these representations with the things-in-themselves [3, p. 103]. Computational definitions of reason typically address mathematical representation and estimation but can involve: a pre-cached knowledge-schema (e.g. slot/terminal assignments, operation procedures); the agent's ontological commitment towards a particular task, story-world, or knowledge domain; a current performance measure progress and evaluation results conforming to a well-defined (and possibly hard-coded) utility function combined with the expected probability-distribution of an incoming percept-stream; and a consistent and persistent semantic interpretation of empirical sense data which can be verified by the scientific process, hard-coded interpretations and the faculty of mathematical estimation [2, 7, 10]. Computational reason mostly draws from procedural and semantic memory stores. Prospective memory storage can also be useful when predicting a new probability distribution. With the third definition in particular, a rational agent's performance measure would also be influenced by competency questions. A reasonable agent must constantly ask new competency questions so that it continually evaluate and update its ontology (imagined world-model) in a stochastic real-time environment [7]. During this query-updating process, the agent might also maintain a reasonable level of competency via maintaining a pointer that continually refers back to its (hopefully) well-defined original utility function and policy. Having competency questions answered with empirical evidence will conserve ontological continuity and deterministic consistency [8]. Regardless of whether these questions were hard-coded by a programmer or self-coded by the agent; rationally optimal questions will already address task-domain procedural, episodic and/or semantic knowledge within the agent's ontology/KB slot terminals. Some of this knowledge is tautological.

Reason in computational terms ranges from mundane activities such as searching in a look-up table to situation calculus and hierarchical planning procedures [10, p. 27]. Reasonable knowledge-queries about the world involve: discrete subjects, predicates, objects, extrinsic rewards and rules as they are not updated continuously. All memory

banks come into play when the agent when addresses "reasonable" questions but the most commonly located memory banks for this process are procedural and semantic. Therefore, this list of competency questions must engage known formulas (e.g. math equations, more constants than variables and [first-order] logic theorems) and peer-reviewed empirical evidence about the worlds the agent wishes to address and become competent within. There are expectations for retrieval-time once these questions are asked by the agent. Optimization algorithms are preferred to ensure latency and memory consumption is minimized. Ultimately, a rational agent can be understood as carrying out an established utility function and a policy with the goal of summing a maximum number of environmental-rewards that are tallied up until the end of the agent's finite lifetime.

2 Kantian Cognitive Intensities from Aesthetic Philosophy

Kantian cognitive intensities are more suitable for addressing AGI mind-component configurations than with his immediate predecessors (e.g. Burke, 1756). Edmund Burke's conception of the sublime intensity, for example, is limited to involving proto-intelligent representations and affective responses based on fear and the agent's need for safety rather than on cognitive interactivity [10]. Kant, conversely, identified cognitive intensity levels from aesthetic philosophy which can also be hermeneutically mapped to the artificial mind-design of rational and deliberative agents. The lowest intensity can be mapped to the least intelligent agent-category while the highest Kantian intensity could only be addressed by a robust AGI-agent [2]. Two of the Kantian intensities (i.e. the good and the strange) correspond less to intellectual capabilities and more to deontic processes (i.e. the former) and solely to aesthetic contemplation (i.e. the latter). For this reason, those two levels will be disregarded from this AGI cognitive capability discussion. The scope will also be limited to focuses on one polarity valence. For example, while the beautiful intensity possesses its inverse valence, the ugly, this awareness of opposite aesthetic valences does not contribute to the overall understanding of how Kantian aesthetic intensity levels can be mapped towards cognitive capability requirements for AGI-agents.

The following Table 1 will therefore show diverse categorical gradations of cognitive intensities that will discuss moderate (agreeable) to the highest intellectual affordances (transcendent sublime). These intensity categories begin to diverge towards AGI intelligence levels once the agent contains and understands computational representations that allow to the agent's mind to act according to more disinterested teleological imperatives rather than single-mindedly through its own immediate self-interest (as stated by the programmer). Only the latter two disinterested intensities apply directly to AGI-level intelligence so these will be focused on the most in this paper. The lowest (first) intensity level has been omitted as it only describes pre-intellectual capabilities for handling rudimentary domestic-level routine tasks.

Table 1. Kantian cognitive (aesthetic) intensities and associated agent-categories

Agent level	Cognitive intensity level	Priorities
Pre-AI, AI	**Mundane**	N/A
AI (state-of-the-art)	**Agreeable**	Self-interest, reason
AI, Proto-AGI, AGI	**Beautiful**	Disinterest, imagination
Up to/including AGI	**Transcendent Sublime**	Imagination > reason

2.1 The Agreeable (State-of-the-Art AI)

This second Kantian level operates at the baseline where most readers would consider an agent to be "intelligent". An agreeable agent will have computational components such as goals, reinforcement signal processors, and crude metacognitive structuring that represent a rudimentary sense of self in order to be rewarded for thinking about and acting within its own world model-specified "self-interest". However, this intensity is primarily sustained in only those agents that exclusively pursue limited self-interested pursuits. Agents that single-mindedly pursue limited interests are conventionally represented as narrow AI rather than as agents more acclimatized to general intellectual contemplation (i.e. AGIs). It is this intensity where an agent will be explicitly programmed to optimize a self-interested balance between computational representations of imagination and reasonable estimation. In an agreeable agent's mind, a superficial processing of semantics and some rudimentary awareness – at least of production rules as well as finite-states and available actions within each state - would come into play (more so than with a mundane reactive mind, at least). This agreeable agent would spend cognitive resources deliberating over what was semantically essential, functionally practical, and/or computationally robust (agreeable) versus that which was semantically irrelevant, cognitively dissonant, and/or computationally taxing (disagreeable). Explicit functional interactivity serves as the rubric for any experience of cognitive interactivity at this level. One identifying feature of this particular intensity when it comes to contemplative purposes is that everything is practically contemplated for the agent's self-interest (i.e. intended functionality). Interestingly, an agreeable agent's mind might not necessarily require an explicitly represented self-concept in order to act in its best "self-interest". This level of contemplation would ensure though that this agent mentally optimizes its intelligence modules and/or processes (i.e. algorithms) in such a way that it can carry out the most functional and practical decisions and/or actions. It is also possible that an agreeable agent would focus on empirical output and related behavioral output, perhaps even with the intent of successfully faking an AGI-agent's ability to function at deeper mentalist aspects of cognition [11, 12]. This level of contemplative intensity exclusively focuses on common-sense and mostly rational functional interactivity even for more imaginative cognitive deliberations. Agreeable agent-minds should not think up (imagine) or contemplate anything that it cannot potentially act upon in a particular world. Additionally, this agreeable agent's mind should not waste computational time-cycles contemplating any more conceivable inputs than what can be immediately perceived from an empirically derivable source. At this intensity level, semantic layers (esp. explicit knowledge types and conceivably alternate inferences) are only

contemplated for functional and practical ends. Overthinking and deep-thinking are discouraged from agent-mind contemplation at this intensity level.

Agreeable Agent Categories. Agent-minds operating at an agreeable level of contemplating intensity would optimally range in intelligence from a model-based reflex agent to a goal-based agent. It is conceivable that a utility-based agent could also engage with this intensity but having an awareness of a utility usually requires a more holistic higher and deeper semantic (extra-functional and extra-practical) understanding of how each of this agent's goals and motivations relate to one another. The more practical and functional the overall utility would be in the larger interactive schema, the more likely an agreeable agent could belong to the utility-based agent category. The state-of-the art currently focuses on narrow agreeable AI and only in recent years, has an AGI community arisen that is interested in designing agents that can contemplate at higher intensities than the agreeable [13].

2.2 The Beautiful (Proto-AGI, Early AGI)

Unlike the agreeable, a higher cognitive engagement results from an AGI-capable agent experiencing the beautiful. This is because no perceivable and/or explicitly programmed personal gain nor loss for the agent would result as a consequence of this contemplation. It is a functionless aesthetic pleasure-in-itself without any concern for meeting intrinsic and extrinsic goals, drives, and motivations. For example, an agent might take a disinterested formal pleasure in contemplating colors, textures, patterns and forms for their own sake regardless of whether they explicitly contribute to the agent's survival and/or programmed drives. The beautiful contemplative level runs at a slightly higher intensity level than the agreeable intensity. Firstly, this intensity type engages in disinterested rather than self-interested contemplation modalities. By entering into a cognitive state-intensity level of contemplative disinterestedness, a beautiful agent is likely to prefer entertaining imagined over reasonable cognitive states without explicit teleological ends in mind. The contemplation of conceivable outcomes requires that the agent's mind can imagine possible inferential outcomes (i.e. using its imagination via simulation) while being able to transcend its own goals, drives, and motivations. The act of deploying an artificial imagination with metacognitive reflection very likely indicates the non-trivial usage of more computational cycles.

Beautiful Agent Categories. This intensity relies on disinterestedness and at the very least, likely requires some computational representation of value that can be used to assess the overall utility of goals, drives, and motivations [14]. It is only through a meta-evaluation of these self-interested goals would a beautiful agent's mind be sufficiently able to transcend each goal's self-interested purpose and view the overall utility from the vantage of cognitive disinterestedness. Therefore, the beautiful mind must possess enough AGI-ready metacognitive capabilities to assess the teleological value of no longer acting in its immediate self-interest. The lowest agent-category for assessing self-interested goals from a disinterested teleological perspective would be that of the utility-based agent. It is just as likely that learning agents [3] and even knowledge-seeking

agents would contemplate percepts at this intensity level [15]. This intensity level would also require an agent's mind to be able to imagine unreasonable scenarios for contemplation purposes while still retaining robust cognitive functionality. Utility-based agents would meet the minimum requirements to consistently imagine concepts that are not always contingent on precise reasonable estimation.

2.3 The Transcendent Sublime (up to and Including AGI)

This particular contemplative (aesthetic) intensity differs from the other ones in that the Kantian sublime transcends dualistic notions of valence. Kant's official definition of the transcendent sublime differs from Burke's sublime in that the cognitive interplay between one's imaginative and reasonable faculties are more important to contemplative (reflective) judgment than to Burke's emotional appeal to the empirical sensations of fear and looming mortality. In particular, the Kantian transcendent sublime *"at once* [,] *a feeling of displeasure, arising from the inadequacy of imagination in the aesthetic estimation of magnitude to attain to its estimation by reason, and a simultaneously awakened pleasure, arising from this very judgment of the inadequacy of sense being in accord with ideas of reason, so far as the effort to attain to these is for us the law"* [1, p. 106]. Immediately after sublime phenomena is present and noticeable, the agent's mind initiates a dynamic race between accelerating cognitive faculties of imagination and reason. Kant's transcendent sublime is a threshold state that is on the absolute edge of surpassing both of these competing faculties and is placed at the highest expected contemplative intensity and reserved for the highest intelligent agent-categories. To have the imagination completely exceed the limits of intellectual functionality is not sublime as the phenomenon would merely be beyond any agent's comprehension (including a super-intelligent agent).

A computational representation of the transcendent sublime can include the uncanny ability to seemingly contemplate and predict a prospective generative percept-stream (future states, operators, probability distributions, actions etc.) and/or simulated input-output procedure that exists primarily outside of or beyond (or in parallel to) the agent's and the virtual-agents' established capabilities of mathematical estimation. The transcendent sublime must include but not surpass an agent's understanding of: a sensory apparatus, and pre-cached knowledge-schema (e.g. slot/terminal assignments, operation procedures). While experiencing this cognitive threshold state, an AGI-agent could also engage in a form of contemplative creative-reinterpretation (or misinterpretation) of a KB and/or ontology that can include sensory data as its imaginary inspiration. Symbols being grounded in an artificially produced transcendent sublime experience might make use of infinite recursion within the code-structure and/or within an AGI-agent's self-improvement mechanism(s). From this definition, an agent's recursive self-improvement would possess capabilities, behaviours, beliefs, and actions that eventually surpass the imagination and mathematical estimation of the original seed programmers [16]. An AGI-agent would also display an excess of initial reasonable and imagined cognitive states, operators, estimations, and decisions that surpass a narrow-AI agent's ability to mathematically estimate the likelihood of successor candidates for those states, operators, search-spaces, and decisions etc. Such computationally operational spaces should

be perceived as vast enough to temporarily confound the mathematical and imaginative estimative predictions of the perceiving agent's mind. However, these operational spaces should also be expressed with contextual appropriateness in order to not have the agent's preferred or selected states/operators appear as arbitrary or random –within its own mind.

Transcendent Sublime Agent Categories. Only a small number of artificially intelligent agent minds can reach the AGI-pertinent benchmark of contemplating the cognitive threshold of the transcendent sublime before either losing symbol-grounding tractability and/or no longer grasping what is being contemplated in a particular world-model (and/or within the artificial agent's own imagination). Even some human BGI-minds would have difficulty being able to cognitively function once imaginative possibilities surpasses the ability to reasonably estimate the plausible outcomes of these possibilities. Ultimately, an artificial agent must be of a sufficiently high enough cognitive category to be able to not only handle cognitive disinhibition but also metacognitive monitoring, reasoning, and regulation [9]. An agent functioning at a lower level than a proto-AGI utility-based agent might simply get confused more easily and/or even ignore the contemplated phenomenon outright for not showing any immediate teleological (i.e. functional and practical) value. The AGI-agent's mind must be able to imagine many different cognitive states in order to explain the probability of alternate semantic explanations for the existence of a particular percept or concept. The state-space of this explanatory inference making would exponentially scale to be barely manageable for most minds and would certainly tax architectural memory systems and decision-making mechanisms. Any agent-level lower than this, and the agent's mind will not be able to even imagine or estimate the cognitive phenomenon worth deeply contemplating.

3 Conclusion – Research Value for AGI

Contemplation and its associated intensification processes are semantically very difficult to articulate when assessing computational rubrics for AGI-agent minds. Kantian aesthetic principles in this semantic context, act as operational metaphors for understanding the hermeneutics of generally intelligent cognitive mechanisms and agent-levels. Operational metaphors – such as the Kantian transcendent sublime as an aesthetic metaphor for the cognitive limits of contemplative intensity - are more than mere rhetorical devices used to loosely describe the semantic ambiguities of a phenomenon in qualitative terms. Within the AI and AGI communities, metaphors alone are sufficient as a "[…] *conceptual lever that allows a system* [incl. agent] *to extend its model of the world"* [17, p. 1]. Through the methodological process of metaphorical creative introspection, one can leverage a trans-disciplinary knowledge base from the humanities domain and transfer this knowledge about aesthetic principles to the more technological domains of theoretical AGI and its associated agent-categories.

References

1. Kant, I.: Critique of Judgment (1790). Pluhar, W.S. [Trans.]. Hackett, Indianapolis (1987)
2. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall, Englewood Cliffs (2010)
3. Russell, S.: Learning agents for uncertain environments. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 101–103. ACM, Madison (1998)
4. Kant, I.: Critique of Pure Reason (1781). Meiklejohn, J.M.D. [Trans.]. Courier Corporation, North Chelmsford (2003)
5. Marques, H.G., Holland, O.: Architectures for functional imagination. Neurocomputing **72**(4–6), 743–759 (2009)
6. Li, J., Laird, J.E.: The computational problem of prospective memory retrieval. In: Proceedings of the 17th International Conference on Cognitive Modeling, Ottawa, Canada, pp. 1–6 (2013)
7. Barbosa Fernandes, P.C., Guizzardi, R.S.S., Guizzardi, G.: Using goal modeling to capture competency questions in ontology-based systems. J. Inf. Data Manag. **2**(3), 527–540 (2011)
8. De Blanc, P.: Ontological crises in artificial agents' value systems. In: The Singularity Institute [The Machine Intelligence Research Institute], San Francisco, CA, pp. 1–7 (2011)
9. Zilberstein, S.: Metareasoning and bounded rationality. In: Cox, M.T., Raja, A. (eds.) Metareasoning: Thinking About Thinking, pp. 27–40. MIT Press, Cambridge (2011)
10. Burke, E.: A Philosophical Enquiry into the Origin of our Ideas of the Sublime and Beautiful (1756). University of Notre Dame Press, South Bend (1968)
11. Uttal, W.R.: The War Between Mentalism and Behaviorism: On the Accessibility of Mental Processes. Psychology Press, Philadelphia (1999)
12. Colton, S.: Creativity versus the perception of creativity in computational systems. In: Proceedings of the AAAI Spring Symposium on Creative Systems, Palo Alto, CA, pp. 1–7 (2008). [14]
13. Goertzel, B., Pennachin, C.: Artificial General Intelligence. Springer, New York (2007)
14. Dewey, D.: Learning what to value. In: Schmidhuber, J., Thórisson, Kristinn R., Looks, M. (eds.) AGI 2011. LNCS (LNAI), vol. 6830, pp. 309–314. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22887-2_35
15. Orseau, L., Ring, M.: Self-modification and mortality in artificial agents. In: Schmidhuber, J., Thórisson, Kristinn R., Looks, M. (eds.) AGI 2011. LNCS (LNAI), vol. 6830, pp. 1–10. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22887-2_1
16. Nivel, E., et al.: Bounded Recursive Self-Improvement. arXiv preprint arXiv:1312.6764 (2013)
17. Veale, T., Li, G.: Creative introspection and knowledge acquisition: learning about the world thru introspective questions and exploratory metaphors. In: Proceedings of the AAAI, San Francisco, CA, pp. 1–7 (2011)
18. Grosso, R.W., et al.: The evolution of Protégé: an environment for knowledge-based systems development. Int. J. Hum. Comput. Stud. **58**(1), 89–123 (2003)