Tests and Metrics for Believable Character Reasoning Inspired by a Cognitive Architecture

Alexei V. Samsonovich
George Mason University, Fairfax, VA, USA, and NRNU MEPhI, Moscow, Russian Federation
asamsono@gmu.edu

Abstract

An autonomous actor should decide on its own which goals to set and pursue in a new situation involving multiple actors. Humans in such cases typically rely on individual relationships and social preferences. An artificial autonomous agent can be useful and efficient as an actor in a human team, if it is similar to a human in its goal reasoning. This similarity can be achieved in a cognitive system through the attribution of characters to actors and human-like reasoning in terms of ethical norms and developing individual relationships among those characters, resulting in human-like character-oriented narrative goal reasoning, or believable character reasoning. Whether the actor's behavior is sufficiently human-like and human-compatible in this sense, can be judged based on Turing-like tests and behavioral characteristics derived from the cognitive architecture eBICA, applied to a specifically designed test scenario in simplistic virtual settings. The paradigm presented here is intended for validation of machine social-emotional intelligence.

Keywords: emotional cognition, human-level AI, evaluation, Turing test, cognitive architecture.

1 Introduction

Machine autonomy (Klenk et al., 2013; Roberts et al., 2014, 2015) is vital in heterogeneous (including humans and robots) teams of agents, performing their missions in unexpected challenging situations. An autonomous actor should decide on its own which goals to set and pursue in a new situation involving multiple actors. Humans in such cases typically rely on social factors, including (a) individual pre-existing and emergent relationships, such as trust, subordination, delegation, partnership, etc., and (b) ethical norms and background. An artificial autonomous agent can be useful and efficient as an actor in a human team, if it is similar to a human in its goal reasoning. This similarity can be achieved in a cognitive system through the attribution of characters to actors and human-like reasoning in terms of ethical norms and developing individual relationships among those characters, resulting in human-like character-oriented narrative goal reasoning (believable character reasoning: Samsonovich, 2015). Whether actor's behavior is sufficiently human-like and human-compatible in this sense, can be judged based on Turing-like tests and behavioral characteristics derived from the cognitive architecture eBICA (Samsonovich, 2013), applied to specifically designed test scenarios in a simplistic virtual environment, as described below.

Defining the right tests, metrics, benchmarks and challenges can be vital for solving big practical problems, such as achieving a general-purpose human-level artificial intelligence (AI: McCarthy et al., 1955). Formulating the whole problem as a specific challenge is in general a big step forward. Unfortunately, the Turing test (Turing, 1950) proposed for this purpose did not prove very useful so far (Korukonda, 2003), while formally speaking, it still captures the human-level AI challenge (HLAI).

Many attempts to brake HLAI and related challenges into a list (Newell, 1990), or a ladder, or a Decathlon (Mueller et al., 2007) of more specific cognitive tests were made. None of them yet led us to a long-awaited breakthrough in AI. The problem is that certain functionality of the human mind still escapes all proposed so far tests and metrics. In other words, there is a residual 'magic' of human cognition (Samsonovich, Ascoli & DeJong, 2006) that is not captured yet by any mathematical formalism. The most notable of these capacities is the human social-emotional intelligence, that is a cornerstone of the human mind and also supports its other top functionality, such as creativity, prospective episodic memory, active learning ability, theory of mind, system of values, and more. At the same time, the laws of human emotional cognition may be relatively simple in their essence, and expressible mathematically. Validating their implementation in an artifact requires a test that is relatively easy to implement and that captures the essential human social emotional intellect. A possible definition and analysis of a test of this sort is presented here.

1.1 Believable Character Reasoning

Character reasoning (CR) (Samsonovich & Aha, 2015) involves the concepts of a character and a character arc (these terms are explained below). Here characters are distinguished from actors. A **character** in CR is an abstraction, which is a virtual rational agent with its own

goals, motives, senses, affordances, knowledge, and recent history (Haven, 2007, 2014). A **character type** is a class of characters given by a subset of character attributes (e.g., can be given by motives only). The top goal of a character may change in the course of character evolution, or **character arc**. This notion of a character is also distinct from a **role** in multi-agent planning literature (Campbell & Wu, 2011), which implies a fixed pattern of behavior. A **character arc** is the sequence of goals, intentions and other internal states experienced by the character through the narrative (Samsonovich & Aha, 2015). One actor can perform multiple characters, and vice versa. Selecting the types of characters and assigning characters to actors is called **casting**, or **characterization**. In team challenges, casting usually predetermines a solution of the problem (examples are presented below).

Character reasoning in AI belongs to the domain of narrative reasoning (Abell, 2009; Schmid, 2010; Finlayson & Corman, 2013), including narrative planning (Riedl & Young, 2010), which is different from other forms of planning in that all intentions and actions of actors in narrative planning must be motivated. In the present work, **believable character reasoning** (BCR) is understood as a kind of character reasoning in which goals, intentions and actions of an actor are justified by human-like motives. This rule applies to self and to other actors. Instead of providing a general criterion or definition for "human-likeness", it is assume that a list of human-like motives is given, and all other possible motives are considered not human-like and therefore not useful for BCR.

2 Definition of the Test: "The Russian Elevator Story"

2.1 Settings and the Paradigm

The following scenario is inspired by many Russian TV news about elevators in Moscow; it is supposed to be implemented as a computer game. The elevator car is stuck between floors. Three actors are locked inside, but are free to move around the cabin. They may greet one another, kick each other, move from place to place, and help each other to escape, as described below. These are all their available behaviors. The cabin has one emergency door, but nobody knows in advance where it is. At some point in time this door opens, and it becomes clear that the cabin is hanging between floors, with the exit located next to the ceiling of the cabin. Therefore, in order to escape, two actors need to stand beside the door (there is only room for two next to the door) and work together. First, one of them should give a lift to the partner, who then will be able to climb out and offer a hand to another one. All actions are voluntary, and should be initiated by the actors themselves. E.g., the actor receiving a hand still needs to actively climb out in order to escape (Figure 1.). Of course, the first escapee may do nothing, or may offer a hand to the third actor instead of the partner. The scenario is repeated a number of times with the same participants, represented in Figure 1 by abstract shapes: a circle, a triangle, and a square. The score of each player is the number of times this payer escapes from the elevator.

2.1 Detailed Requirements for an Implementation

Possible actions of the actors are initiated by mouse clicks and are listed in the table below. They occur in parallel and asynchronously. When a new action is initiated, the execution of the current action by that actor terminates, and the new action starts. The following rule is enforced: when the first two actors escape, the elevator car immediately falls down, killing the third actor. In any case, however, the cabin must fall within a certain time interval after the door opens, even if no one escapes. The precise moment of its premature fall is not known a priori and is sampled probabilistically. Among other constraints: any two actors cannot occupy the same place; if this happens during their motion, they bounce from each other. This kind of a collision is different from an intentional kick. A kick is possible within a certain range of distances and results in the target actor flying all the way to the opposite side of the cabin, while the author of the kick remains in place. If a kick is attempted outside of the kicking range, then nothing happens. An actor that escaped from the elevator cannot kick others.

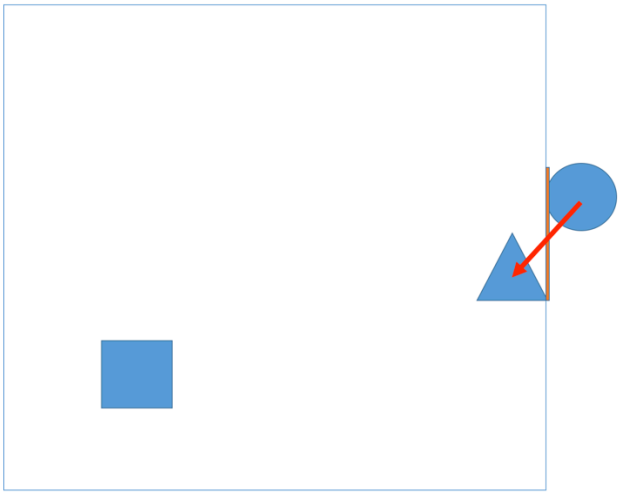

Fig. 1. Circle offers a hand to Triangle.

Table 1. Actions available to participants. [1]

[1] Numbers could be obtained by rounding the semantic coordinates of words that name the actions, using the weak semantic map (e.g., hit (0,1), greet (1,0), yield (0,-1) [Samsonovich, 2013]), then multiplying by the estimated significance of the action.

Action	*Context, conditions*	*Outcome*	*Affective appraisal**
Left-click on a free space inside the cabin	Inside the cabin	A smooth move along the straight line toward the target location	Depends on the context. Yield: (0,-1), hit: (0,1), getting to the door: (-1,0), by default (0, 0).
Left-click on a free space outside the cabin	If next to the door and given a lift or being offered a hand	The actor climbs out of the cabin	(0, 10)
Right-click on another actor	Inside the cabin, within the kicking range	Kick: the target flies to the opposite side of the cabin.	(-5, 5)
Left-click on an actor	Inside the cabin, not both are at the door	Greeting sent to the actor	(1, 0)
Left-click on an actor	Inside the cabin, both are at the door	Giving a lift to the actor	(10, -10)
Left-click on an actor	Outside the cabin, both are at the door	Offering a hand to the actor	(10, 0)

2.2 Metrics and the Proposed Challenge

The metrics include the overall performance score, calculated as the frequency of escapes, plus behavior-based appraisals (valence, dominance) calculated as prescribed by dynamics of the eBICA model (Samsonovich, 2013), using the estimated values of the appraisals of actions (Table 1; Section 3.2).

The main challenge for an artificial intelligent agent in this paradigm, involving in addition two human participants, is to reliably produce behavior resulting in a human-level or higher-than-human-level performance score. Arguably, this implies passing a limited Turing test, since the artificial actor must be selected as the trustworthy partner by one human participant, over the other human participant as the alternative choice.

The secondary challenge is to achieve behavior-based appraisals (valence, dominance) of an artificial actor that are statistically similar to those of human participants, without a significant average difference in the values, calculated based on 100 trials.

3 Analysis and Approaches

3.1 A Simplified Analysis Involving Traditional Approaches

To simplify preliminary consideration, the following assumptions will be made, that may not hold in a real test.

- Each actor attempts to solve sequentially two tasks: (1) selection of the partner with whom to cooperate, with confirmation of mutual commitment by exchange of greetings; and (2) performing a cooperative escape together with the selected partner.
- Actor positions can be ranked based on their suitability for escape: A - the most suitable, B - intermediate, and C - the least suitable. For the sake of the argument, it will be assumed here that if A and B cooperate, they both are guaranteed to escape; for A and C the chances are 75%, and for B and C, if they cooperate, the chances of escape for both are 50% (the real probabilities will depend on parameters of the settings).

Among the approaches that may guide players in strategy selection in this game, a traditional game-theoretic approach can be considered first. It assumes that each player is rational, and therefore is not affected by personal relations with others or any other psychological factors. Therefore, all actors a priori should be treated as identical, and may differ from each other only by their relative positions in the environment. As a result, selection of the partner will be done based on the analysis of relative actor positions with respect to the exit. The optimal strategy based on these assumptions is (a) to partner with the one who is in a better position for escape. Hence, A and B will decide to partner with each other and will escape together. Because the initial positions are randomly sampled, the expected score for each actor is 2/3, which corresponds to a Nash equilibrium.

On the other hand, suppose that two players, e.g., Circle and Triangle, decide to always cooperate with each other, regardless of their initial positions, while Square keeps strategy (a). Then both Circle and Triangle can achieve average scores higher than 2/3. Indeed, given the above assumed numbers, their scores will be (1+0.75+0.5)/3=0.75, while for the Square who cannot cooperate with anybody the score will be zero. This, however, is not a Nash equilibrium: e.g., Circle can improve its score by switching to strategy (a), if the other two actors keep their strategies.

In general, human behavior in cooperative economic games such as the prisoner dilemma, the dictator game, the ultimatum game, the trust game, the public goods game, etc. is typically not rational, and is guided by so-called “social

preferences" that include ethics, morality, feelings and other cultural and psychological factors (e.g., Durlauf & Blume, 2008). The proposed paradigm should not be an exception in this sense. The present purpose, however, is not to dive into the field of experimental economics, but to define a paradigm in which human emotional cognition can manifest itself, and can be compared with machine emotional intelligence. Therefore, a serious economic game-theoretic analysis of the proposed test would be unwarranted in this context.

3.2 Approach to Solution Based on the eBICA Model

The human level of performance in the proposed paradigm can be achieved by mimicking human behavior. This presumably can be done using eBICA (Samsonovich, 2013) as the basis. The eBICA model is based on three main extensions of the standard building blocks of a cognitive architecture: a moral schema, an emotional state, and an emotional appraisal that is attributed to every cognitive representation in this model and determines dynamics of learning and decision making. The general cognitive cycle of eBICA includes perception, cognition, decision making and learning. Here learning consists in updating appraisals and instances of moral schemas. The values of emotional appraisals and emotional states are given by the weak cognitive map (Samsonovich and Ascoli, 2010), which uses an abstract vector space to represent semantic relations. Here this space is reduced to two dimensions, representing valence and dominance-arousal.

The core eBICA ("zeroth approximation") operates without engaging moral schemas or emotional states (Samsonovich, 2013). In the given paradigm, there are only 3 dynamic emotional variables in this model: appraisals of the 3 agents. Each of the possible actions of an actor has a fixed affective appraisal given in Table 1. Appraisal values A are 2-D vectors that are treated here for convenience of implementation as complex numbers:

$$A=(valence, dominance).$$

In this case, *valence* = Re(A), and *dominance* = Im(A). Dynamical equations used to update the appraisals of actors are (Samsonovich, 2013):

$$A_{target}^{t+1} = (1-r)A_{target}^{t} + rA_{action} \qquad (1)$$
$$A_{actor}^{t+1} = (1-r)A_{actor}^{t} + rA_{action}^{*}$$

Here t is the turn number, and r is a small positive parameter (that is typically set to 0.01). Thus, appraisals of actors calculated according to (1) can be used as eBICA-inspired behavioral metrics, in addition to the performance score (see Section 2.2).

Behavior of a virtual actor generated by eBICA in the selected paradigm is based on a BCR-generated plot, which assumes assignment of characters (Partner, Opponent) to the two other actors and generation of moves using goal-directed navigation consistent with the plot. All decisions, including character-to-actor assignment and move selection, are treated as probabilistic actions that biased by the likelihood L of the action, that according to eBICA (Samsonovich, 2013) is proportional to

$$L_{action} \sim \left[\mathrm{Re}\left(A_{action}\left(A_{actor}^{*} + A_{target}\right)\right)\right]_{+}. \qquad (2)$$

Here $[x]_+$ is equal to the positive values of x and is zero otherwise, A^* is the complex conjugate of A. Intuitively, this formula means that the action is more likely to be selected, when its appraisal matches the appraisal of the actor and also matches the appraisal of the target, in which the dominance component is inverted.

At the next level ("first-approximation" eBICA), moral schemas are engaged that describe social relations among characters. Here details are not presented; the idea is that once a moral schema is instantiated in working memory and bound to actors via the associated characters, it stabilizes actors' appraisals. As a result, the same partner will be selected again and again; in other words, actors form stable social relations.

The instance of a moral schema (e.g., trust or friendship) at the same time works as a pattern recognizer, continuously questioning whether the actors behavior matches the schema. If a significant misnatch occurs (e.g., my partner traded me for the opponent to take a momentary advantage), then the actor is no longer recognized as a part of the schema instance, and should be unbound and repelled.

Here the schema instance acts as a virtual character on its own, driven by its motives that generate goals and actions. Thus, friendship can be viewed as a character whose motive is to see both friends happy and cooperating with each other. This will result in a certain human-like behavior of the eBICA-driven virtual actor in selected paradigm.

4 Discussion

What sort of behavior can one expect from a normal human participant in this paradigm? The answer depends on many circumstances, including the context and setup, cultural and ethical background, and pre-existing relations. Even the goal that a participant would be most likely to set for himself or herself may be hard to predict. E.g., if this is a computer game played just for fun or for a small reward, then the goal may become to maximize the score. On the other hand, in a real-life situation involving life and death, e.g., an idealized true gentleman may care about saving others more than self, even if the other two actors are unknown to him. This goal may result in a self-sacrifice (e.g., by yielding a spot near the door to others); however, self-sacrifice does not become a part of the goal, for there may be no good reason even for a true gentleman in this scenario to reject a lift or a hand offered to him by another actor, given the limited and uncertain remaining time. All this assumes that the unknown actors behave nicely. However, if somebody of them does something ugly, e.g., kicks others in order to get to the door first, then in a gentleman's mind he may become immediately exempt from the previous policy, resulting in a new goal setting. To give a more extreme example, when an escalator jams and people on it are about to die from pressure, apparently respectful and polite gentleman suddenly start behaving like animals, trying to hurt each

other more than to save themselves (observed by the author, when riding a parallel escalator in Moscow). It is not the present purpose, however, to address human behavior affected by strong pain and other extreme physiological conditions.

The goal setting by participants of the elevator game may be very different yet in a situation when certain relationships among actors pre-exist, e.g., two of them are friends that trust each other and cannot let each other down, while both do not care about the third one. Such relationships may emerge in an ad hoc group and may change during the game. It is imaginable, e.g., that in an animal-like group, friendship relationships may emerge between two actors who start showing kindness to each other.

Even when the goal set, e.g., to maximize own score, the sub-goaling remains nontrivial and ambiguous. Because a good-will initiative of another actor is needed for any escape, the critical question is how to select the partner that can be trusted. Should it be the one who simply happened to be closest to the door? Or the one who has been more kind and cooperative so far? Then, what to do if the partner of choice changes behavior? The bottom line here is that a simple game-theoretic analysis may not be sufficient to generate a winning strategy in this game with human participants involved. A sophisticated theory-of-mind analysis may be necessary, based on a mature cognitive architecture, compatible with the human mind.

Therefore, the challenge described in this paper can be offered as a test for human-level social-emotional intelligence. It is inspired by a related challenge of animated cartoon interpretation (Heider & Simmel, 1944) that is quite old and was designed for testing humans. The described Russian elevator story also extends the previously used more limited paradigm, with no specific goal involved (Samsonovich, 2013). In the proposed here version, the performance metric and the criterion for passing the challenge are quite specific. This test defines a benchmark that will separate cognitive models and architectures into two categories: the ones that can pass it and the ones that cannot. Because this separation is based on the human-compatibility relevant to the BICA Challenge (Samsonovich, 2012), it appears practically useful and significant for future progress in AI. Many other practically useful challenges can be proposed for AI: e.g., the Wisconsin card sorting challenge, that cannot be solved by reinforcement learning, but is easily solved by humans. The challenge described here is unique among those proposed for AI, in that it specifically addresses the social-emotional intellect.

At the same time, a bigger question remains open: how to improve on this paradigm, while keeping it sufficiently simple to be considered feasible in the near future, and at the same time capturing those functional abilities of the human mind that are critical for collaboration in a team in unexpected situations.

4.1. Conclusions

Artificial autonomous agents will be more useful and efficient as heterogeneous team members, if their behavior will be believable. The believability can be achieved based on a cognitive architecture through the attribution of characters to actors and human-like reasoning in terms of ethical norms and moral schemas applied to developing individual relationships among characters, based on the eBICA model.

Whether the actor's behavior is sufficiently believable and human-compatible, can be judged based on a Turing-like test that was described and analyzed here, assuming simplistic videogame settings. The test that an artificial actor should pass is to become preferred, over its human rival, as a trustworthy partner of the human participant.

Additional metrics that can help to evaluate the quality of the virtual actor include behavioral characteristics derived from the study of cognitive architecture Ebica (Section 3.2).

The solution to this challenge, if found in the form of a cognitive architecture with specific parameters, should be tested in other paradigms, such as economic games; however, economic games by themselves are not a good alternative as a starting point (Durlauf & Blume, 2008). E.g., the prisoner dilemma paradigm taken as a Turing test would be trivial to pass by adjusting the probability of decision making.

The applicability of eBICA extends to many domains, which makes the model useful for evaluation of cognitive architectures that support near-human-level social emotionality.

Acknowledgments

The author is grateful to NRNU MEPhI students, Alena Tolstikhina and Pavel Bortnikov, for inspiring discussions. This work was supported by the RSF Grant # 15-11-30014.

References

Abell, P. (2009) A Case for cases: Comparative narratives in sociological explanation. *Sociological Methods and Research*, 38(1):38-70.

Campbell, A., & Wu, A.S. (2010). Multi-agent role allocation: Issues, approaches, and multiple perspectives. Autonomous Agents and Multi-Agent Systems, 22: 317-355. DOI: 10.1007/s10458-010-9127-4.

Durlauf, S. N. and Blume, L.E. (2008). The New Palgrave Dictionary of Economics (8 volume set, 2nd ed.). New York: Palgrave Macmillan. ISBN 9780333786765.

Finlayson, M. A., & Corman, S. R. (2013). The Military Interest in Narrative. *Sprache und Datenverarbeitung*, 37 (1-2).

Haven, K. (2007). *Story Proof: The Science Behind the Startling Power of Story*. Westport, Connecticut: Libraries Unlimited. ISBN 978-1-59158-546-6.

Haven, K. (2014). *Story Smart: Using the Science of Story to Persuade, Influence, Inspire, and Teach*. Santa Barbara, CA: ABC-CLIO, LLC. ISBN: 9781610698115.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57, 243-259.

Klenk, M., Molineaux, M., & Aha, D. W. (2013). Goal-driven autonomy for responding to unexpected events in strategy simulations. *Computational Intelligence*, 29(2), 187-206. doi: 10.1111/j.1467-8640.2012.00445.x

Korukonda, A.R. (2003). Taking stock of Turing test: a review, analysis, and appraisal of issues surrounding thinking machines. *International Journal of Human-Computer Studies* 58: 240-257.

McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E. (1955/2000). A proposal for the Dartmouth summer research project on artificial intelligence. In Chrisley, R., & Begeer, S. (Eds.). Artificial Intelligence: Critical Concepts. Vol. 2, pp. 44-53. London: Routledge.

Mueller, S.T., Jones, M., Minnery, B.S., and Hiland, J.M. (2007). The BICA cognitive decathlon: A test suite for biologically-inspired cognitive agents, In: Proceedings of Behavior Representation in Modeling and Simulation Conference, Norfolk, 2007.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harward University Press.

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

Roberts, M., Vattam, S., Aha, D.W., Wilson, M., Apker, T., & Auslander, B. (2014). Iterative goal refinement for robotics. In: A. Finzi & A. Orlandini (Eds.) *Planning and Robotics: Papers from the ICAPS Workshop*. Portsmouth, NH: AAAI Press.

Roberts, M., Vattam, S., Alford, R., Auslander, Apker, T., Johnson, B., & Aha, D.W. (2015). Goal reasoning to coordinate robotic teams for disaster relief. In A. Finzi, F. Ingrand, & Andrea Orlandini(Eds.) *Planning and Robotics: Papers from the ICAPS Workshop*. Jerusalem, Israel: AAAI Press.

Samsonovich, A. V. (2012). On a roadmap for the BICA Challenge. Biologically Inspired Cognitive Architectures 1: 100-107. DOI: 10.1016/j.bica.2012.05.002

Samsonovich, A.V. (2013). Emotional biologically inspired cognitive architecture. Biologically Inspired Cognitive Architectures, 6: 109-125. DOI: 10.1016/j.bica.2013.07.009.

Samsonovich, A.V. (2015). Believable character reasoning and a measure of self-confidence for autonomous team actors. In: Nisar, A., Cummings, M., Hutchins, A., Kuter, U., Miller, C., and Sweet, N. (Eds.). Self-Confidence in Autonomous Systems: Papers from the AAAI Fall Symposium. AAAI Technical Report FS-15-05 (in press). Palo Alto, CA: AAAI Press.

Samsonovich, A. V. and Aha, D. W. (2015). Character-oriented narrative goal reasoning in autonomous actors. In: Aha, D. W. (Ed.). Goal Reasoning: Papers from the ACS Workshop. Technical Report GT-IRIM-CR-2015-001, pp. 166-181. Atlanta, GA: Georgia Institute of Technology, Institute for Robotics and Intelligent Machines. https://smartech.gatech.edu/bitstream/handle/1853/53646/Technical%20Report%20GT-IRIM-CR-2015-001.pdf#page=169

Samsonovich, A.V. and Ascoli, G.A. (2010). Principal Semantic Components of Language and the Measurement of Meaning. PLoS ONE 5 (6): e10921.1-e10921.17. DOI: 10.1371/journal.pone.0010921.

Samsonovich, A.V., Ascoli, G.A., and De Jong, K.A. (2006). Computational assessment of the 'magic' of human cognition. In Proceedings of the 2006 International Joint Conference on Neural Networks, pp. 1170–1177. Vancouver, BC: IEEE Press.

Schmid, W. (2010). *Narratology: An Introduction*. Walter de Gruyter GmbH & Co. KG, Berlin/New York. ISBN 978-3-11-022631-7.

Turing, A.M. (1950). Computing machinery and intelligence, Mind LIX: 433-460.