Resource-Constrained Social Evidence Based Cognitive Model for Empathy-Driven Artificial Intelligence

Anton Kolonin[1,2(✉)]

[1] Aigents Group, Novosibirsk, Russia
akolonin@gmail.com
[2] SingularityNET Foundation, Amsterdam, Netherlands

Abstract. Working model of social aspects of human and non-human intelligence is required for social embodiment of artificial general intelligence systems to explain, predict and manage behavioral patterns in multi-agent communities. For this purpose, we propose implementation of resource-constrained social evidence based model and discuss possible implications of its application.

Keywords: Artificial psychology · Artificial general intelligence · Compassion
Cognitive model · Empathy · Social evidence · Social proof

1 Introduction

For complete embodiment of any artificial general intelligence (AGI) system [1], we anticipate the need for social embodiment. That is, besides physical or virtual connections to the world, supplying self-reinforcement feedback, we assume there is a need for social connections supporting social reflections, based on empathy and compassion mutually expressed between members of human or non-human community of natural or artificial beings. In fact, there is known evidence in mass psychology regarding effects that social patterns have on behavior of individuals [2]. These effects may have constructive or destructive implications, depending on the case [3], including such negative scenarios as social engineering and psychological operations [4]. The former may be employed to implement soft control of entire community improving its performance while the latter may be used to abuse and destroy the community. At the same time, need for human-friendly AGI requires comprehension of human values on behalf of AGI system, while these values might get learned in course of self-reinforced co-development of AGI system with humans that it is supposed to serve to. For this purpose, having the system possessing cognitive model capable to learn values of its social environment appears very important.

Earlier works in the area of artificial psychology (AP) involving mathematical modeling of social interactions phenomena and dispute resolution in communities have been carried out by Lefebvre [5]. As it has been suggested by Goertzel and other authors (Kolonin, Pressing, Pennachin) in 2000, basis of social motives of an artificial agent behavior can be grounded on principle of compassion between interacting agents. The

M. Iklé et al. (Eds.): AGI 2018, LNAI 10999, pp. 100–108, 2018.
https://doi.org/10.1007/978-3-319-97676-1_10

definition of the same principle can be called "empathic computing" [6] and applied for study of effects of behavioral modifications in human, non-human and hybrid environments.

In the further discussion, we will be relying on the principles of empathy and compassion as built-in qualities of AGI system, following definition of intelligence made by Goertzel [1] as ability to reach complex goals in complex environments using limited resources. It will be assumed that decision making process of a system capable for social behavior based on these principles can be implemented with fuzzy or probabilistic logic operating with networks or graphs of concepts and relationships [7, 8]. Specifically, we will discuss extensions and implications of the social evidence based cognitive model constrained by resources [9, 10].

In such model, social evidence based decision making process implies that an agents reaches internal consensus in its internal system of reflections of its social referees, being limited by time and amount of power to make these decision timely not over-consuming available energy [9]. In social psychology studies this principle and its implications are well backed up with known phenomenology and notion of social proof identified by Cialdini [2].

The end goal of the work is to engineer working AGI agent capable for empathic and compassionate behavior serving its human environment [11].

2 Model

The suggested resource-constrained social evidence based cognitive model of an AGI agent, based on earlier works [9, 10], assumes the scope of knowledge is represented with atoms [1] being concepts or relationships, with each atom having its truth value, or subjective agent's expression of truth value attached to it.

The scope of atoms may be representing hyper-graph [1], consisting of few segments, as Fig. 1 shows, such as foundation graph, social graph, evidence graph and imagination graph [9]. For further analysis and discussion, we provide following definitions of the segments and their functional relationships.

Foundation graph contains trusted "hardwired" knowledge which does not need fuzzy logic or probabilistic reasoning to infer truth values of knowledge atoms it it. Each i of the atoms F_i represents part of consistent belief system of the knowledge owner, so subjective expression of truth value F_i is fixed maximum value, such as *1.0* in case of reasoning on scale between *0.0* and *1.0*.

Social graph contains weighted relationships with social referees of the knowledge owner, with expression values indicating cumulative level of trust, empathy and compassion in respect to every member j of society S_j. It should be noted that this level may be computed from the rest of the other relationships P_{ij} connecting social referee j to atoms in foundation graph and representing reflection of referee in the belief of the knowledge owner. The latter can be thought as social binding reflecting proximity of of social referee to the knowledge owner self in its own view.

Evidence graph contains facts k of everyday evidence E_k owner of knowledge is being exposed to, with each fact having its expression value. Each of the facts may have

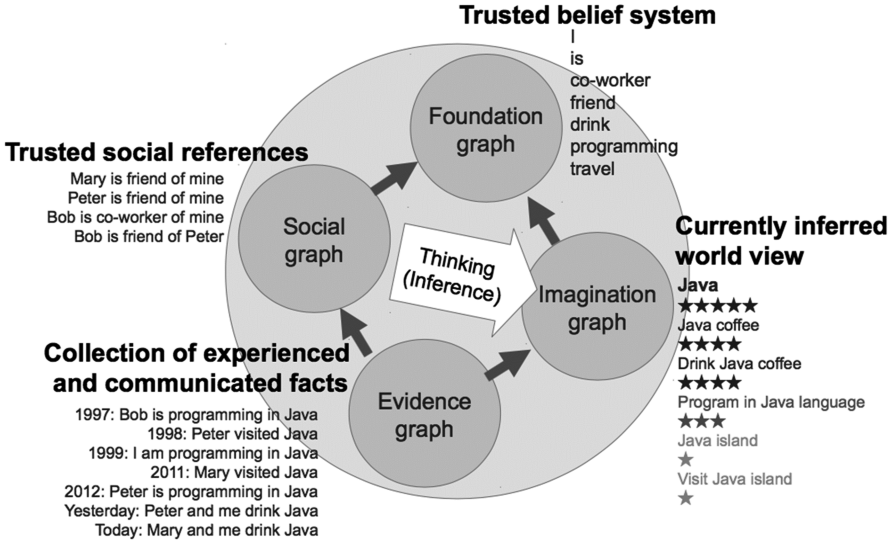

Fig. 1. Social evidence based cognitive model – segmentation of scope of knowledge.

connections Q_{ik} to atoms in foundation graph with expression values representing extent to which the fact is grounded in the belief of the knowledge owner. Also, each of the facts may have connections R_{jk} to members of social graph with expression values indicating extent to which social referee j is though to be responsible for authoring these facts themselves or treating them as reliable and relevant.

Imagination graph contains current view image G_k of the world supplied by evidence facts k to to the extent the facts are grounded in the belief of the owner Q_{ik}, with account to valuation of the facts by referees R_{jk}, including valuations of the referees S_j themselves. The following formula may be used to approximate this dependency.

$$G_k = E_k * (\Sigma_i(Q_{ik}) * \Sigma_j(R_{ik} * S_j)), \qquad S_j = \Sigma_i(P_{ij} * F_i)$$

Obvious interpretation of the formula suggests that expression of fact k in world view image of a subject depends on everything variable above, with growth of it with amount of the raw fact evidence E_k, extent to which is is grounded in core belief system Q_{ik}, and supporting social evidence R_{jk} weighted by social binding S_j.

Further, the idealistic framework described above may be complicated by physical limits on any of the segments, restricting their capacity, so only the atoms with highest degree of expression are retained while the others may by pushed out from agent memory, with different effects in respect to short-term or working memory and long-term one [9, 10]. In simple form, for each segment of the graph, it may be represented with filtering functions F, S, E, and G retaining only top expressed atoms in foundation, social, evidence and imagination graphs, respectively.

$$G_k = G(E(E_k) * \Sigma_i(Q_{ik}) * \Sigma_j(R_{ik} * S(S_j))), \qquad S_j = \Sigma_i(P_{ij} * F(F_i))$$

Important part of the model is that atoms in foundation graph are considered trusted unconditionally, so F_i is always true, while atoms in imagination graph require resource-consuming inference as described above. For atoms G_k that are always true or close to that, the inference makes no sense so resources can be preserved "hardwiring" them into belief system, moving knowledge from segment G to F. For atoms G_k that are not close to true, still requiring inference yet occupying imagination graph often enough to impact on resource consumption by inference, another ways to preserve energy and space are possible. It can be solved with adjusting any of the other variables so that inferred truth values approach to true and respective atoms may follow the scenario above due to increased expression of truth value. Alternatively, the other variables can be adjusted do the truth values get below the filtering functions and respective facts are not involved in the inference at all.

Justification of the model can be considered from few different perspectives. First, there is separation of the scope of knowledge, underlying decision making, into "absolute truths" and "context-specific truths". In hardware and software, the former is more like hardcoded OS-level code operating efficiently in pre-allocated portion of memory while the latter is more like loadable and overloadable applications, operating on top of the former in remaining memory and being swapped optionally. In humans, the former goes to implementation of unconditional stimuli and long-term conditional stimuli associated with deep beliefs such as religion or attachment to liberal or conservative points of view, while the latter corresponds to short-term conditional stimuli and may me changed based on specific circumstances and current mood.

The other justification to split store of information and cognitive processes into segments such as foundation graph and imagination graph is implied by need to provide fast and computationally cheap responses within restricted amount of resources and limited time in respect to operations that are repeating often enough, so they should not consume too much energy, or are critical for survival, so they should be handled rapidly. On the opposite side, events not happening often and not critical to survival may deserver careful consideration within wider context involved different possible inference paths and options. This is like move to rescue children from wild animal or moving car is something fundamental for average human and happens almost unconditionally given core belief, while rescue children from the rain main may be opted out if the rain is warm and children are enjoying the natural shower. It worth noticing that our model assumes the knowledge and cognitive activities that involves it may be moved across these segments due to long-term changes of contexts, as long as environment changes during the life time.

The need for social graph used for social referencing can be justified to keep weights of particular social referees involved in the inference process. It may benefit decision making process introducing social evidence ("social proof" by Cialdini [2]) in cases when there is no sufficient personal evidence to make decision or when there are conflicting personal evidences to be resolved. Since the social evidences from different sources may involve even more conflicting evidences, there is the need to ranking of the sources of evidence by social proximity, expressed in terms of belief proximity. Notable, since either human or artificial being may have no access to internal belief of its peer, we may consider measure of apparent belief or peer's interaction partner to be considered. In humans, for

surrogate measure of social proximity natural and behavioral traits are considered. In artificial beings operating within communities based on open protocols, actual measure of true belief systems may be computed.

3 Analysis

Earlier qualitative empirical modeling of practical implications of the model above are presenting different cases of social engineering [9] as well as overall social dynamics at large scale [10]. Below we discuss how the model works in greater details.

On the left side of the Fig. 2 there is initial state of multi-agent interaction history, where two agents on the top share knowledge atoms A,B with each other, whereas two agents at the bottom share knowledge X,Y. Also, there is agent in the middle sharing A,B with upper ones and X with lower ones, plus it has atom being C communicated to everyone. Finally, the agent on the left at the bottom has Z being communicated with its close circles. Assuming the possessed knowledge resides in foundation graph, due to overlaps in beliefs, the three agents at the top are somewhat closer one to another while the two agents at the bottom are close to each other but distant from the upper three. Respectively, due to different strengths of social bindings, expression of communicated knowledge atoms C and Z is different for different agents, as shown by thickness of arrows representing agent-to-agent interactions.

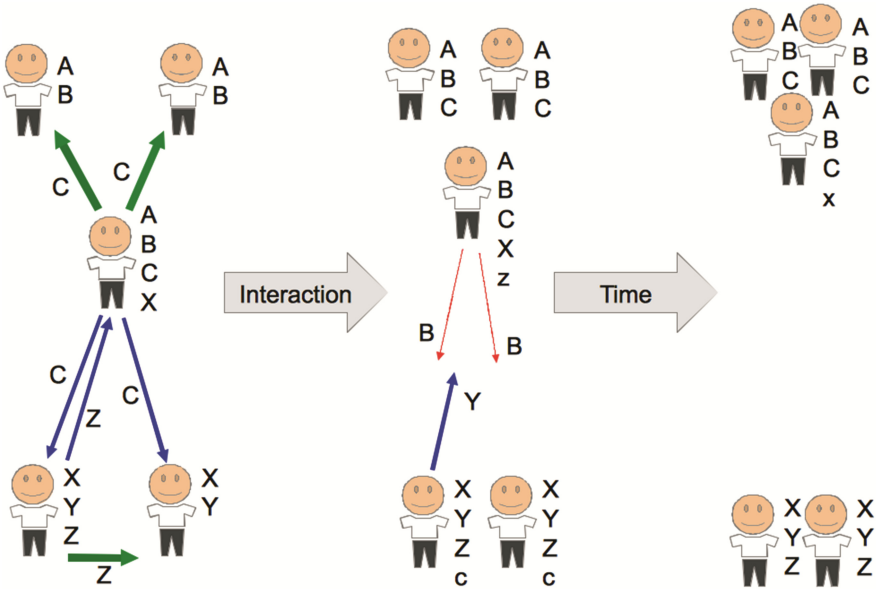

Fig. 2. Explanation of social dynamics due to interaction with knowledge comprehension based on social evidence with impact of limited resources.

The effect of such interaction is shown in the center. Two agents on the top have obtained high expression of C which may be eventually moved to their foundation graph, while two agents at the bottom are given weak expression of C which may stay remaining in imagination graph. Similarly, agent in the middle has got weakened expression of Z. Still, agent on the right at the bottom has got well expressed Z to get retained in its belief. The effect of the interaction is that three agents at the top become socially closer to each other – same as two agents at the bottom, with both groups moved socially further away each from the other. Now, when agent in the middle tries to communicate B to agents at the bottom, their similarity may be not sufficient to let B even entering imagination graph of the latter agents, so the evidence may get completely ignored. Still, Y communicated to middle agent from the bottom may enter its imagination graph, still being weak enough to enter belief system of it.

Over the time, forgetting may take place so the knowledge atoms with weak expression are removed from imagination graphs, so Z may get forgotten by agent in the middle while C may be get forgotten by agents at the bottom. It ends up with further separation of agent society into isolated groups barely sharing any common values.

Social dynamics above justifies earlier discussion [10] that any inhomogeneous community, given no extra input outside affecting expression of some common values, tends to get separated into isolated social clusters eventually. In turn, with external inputs affecting such common values, like mutual benefits or shared existential threats, society can be rather united. On the other hand, having particular inhomogeneous input fed from outside, some of it may be consumed by one part of society but not the other, so in such case internal divergence of society can be even enforced. Practically, the latter effect is being exploited in so called "psychological operations" [4], implementing methods "of social engineering" based on social proof [2]. Respectively, understanding of this dynamics allows to engineer measures to resist psychological operations or social engineering on behalf of society of either artificial agents or humans being subject of such attack vector.

4 Implications and Applications

While numerical simulations based on the model discussed have not been performed yet, qualitative analysis of the model behavior are well confirmed with both positive and negative phenomena found in literature on mass psychology, mathematical modeling and live experiments with social networks [2, 3, 5]. In particular, "social proof" described bu Cialdini [2], methods of directing human masses [3], methods for quantitive modulation of human mood [4] may be turned for good as well as for bad, based on the means and those who applies such method and for which purpose. Within the Aigents project [11, 12], we are trying to build artificial agent compassionate to its human master and its close social environment, so we anticipate what agent learns from its master can not be turned into evil. So, far, current implementation of news monitoring and information extraction agents based on the model learns web surfing preferences and information extraction patterns from the user owner as well as from user's connection in social media, considered as social peer, with proximity of relations between the user and

the peer taken into account [13]. News relevance assessment based on so called "personal relevance" and "social relevance" performed by Aigent can be explored on https://aigents.com website, as shown on Fig. 3.

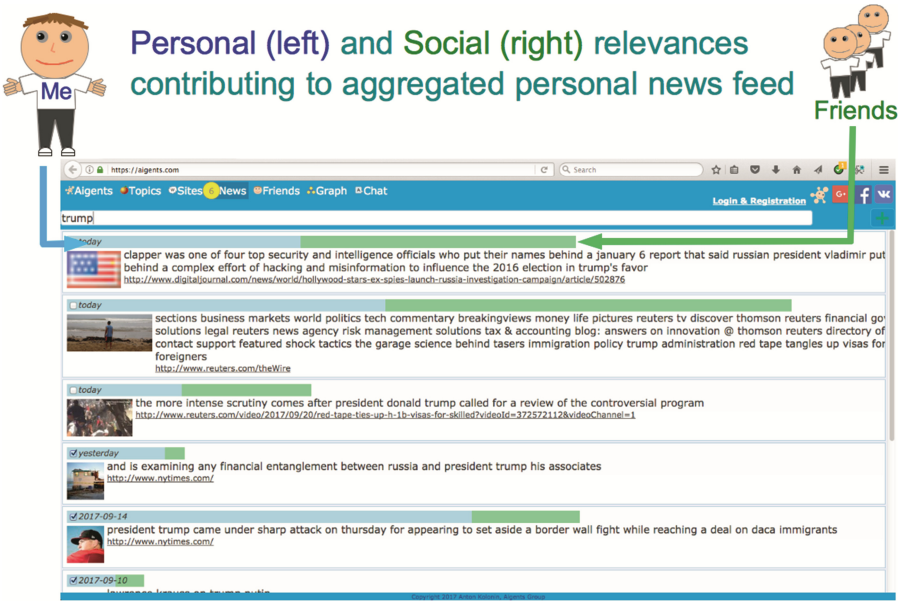

Fig. 3. Personal and social relevances used by news monitoring and information extraction agent. Width of left part of the bar above each of the news items corresponds to the personal relevance, based on experiential learning in course of interaction between the agent and its human owner. Width of right part of the bar corresponds to social relevance learned in the course of interactions with other users, with account to proximity between the users' profiles.

For wider area of applications, we assume that building human-friendly agent of artificial general intelligence can not be rule based but should rather apply reasoning in respect to what can be though as friendly to particular human at the moment and what should be not. Obviously, this reasoning should be efficient, so definition of general intelligence made by Goertzel as "ability to reach complex goals in complex environments giving limited resources" [1] would apply. Hence, different segments of agent memory used by agent for operations on different kinds of data with different performance and efficiency, being adaptable to changing social context appears reasonable for implementation of generic-purpose AGI agents other than just personal assistants specialized for news monitoring and information extraction.

5 Conclusion

We conclude that suggested model is well justified with known phenomenological evidence in the area of mass psychology and may be suited to model behavior of artificial

societies of multi-agent systems, as well as human communities and hybrid human-computer societies. This makes it possible to quantify and predict the resistance to social engineering and psychological operations, and also to model constructive manipulations in respect to target communities. This also provides framework to build cognitive models of AGI creatures capable for human like behavior grounded in empathy end compassion, with possibility of tuning parameters or such cognitive models in course of self-reinforcing interactions.

Still, qualitative empirical modeling and verification by means of phenomenological evidence does not seem sufficient enough to justify the suggested model completely, so simulation modeling of multi-agent societies employing the model is required in the future. The other part of our plan is to implement such model in AGI agent serving to human user as intelligent assistant in the course of interaction with online and social media [11], with prototype now available at https://aigents.com website.

Acknowledgements. This work was inspired by earlier ideas of Ben Goertzel, Jeff Pressing, Cassio Pennachin and Pei Wang in the course of Webmind project targeted to build artificial psyche in 1998–2001.

References

1. Goertzel, B.: CogPrime: An Integrative Architecture for Embodied Artificial General Intelligence. OpenCog, Paris, October 2, 2012 (2012)
2. Cialdini, R.: Influence: The Psychology of Persuasion. ISBN 0-688-12816-5 (1984)
3. Kramer, A., Guillory, J., Hancock, J.: Experimental evidence of massive-scale emotional contagion through social networks. PNAS 2014 **111**(24), 8788–8790 (2014)
4. Nazaretyan, A.: Psychology of mass behavior. ISBN 5-9292-0033-5, PER SE (2001)
5. Lefebvre, V.: Algebra of conscience. Springer, New York (2001). https://doi.org/10.1007/978-94-017-0691-9
6. Nguyen, H., Masthoff, J.: Designing empathic computers: the effect of multimodal empathic feedback using animated agent. In: Proceeding Persuasive '09 Proceedings of the 4th International Conference on Persuasive Technology, Article No. 7, Claremont, California, USA 26–29 April 2009
7. Iklé, M.: Probabilistic Logic Networks in a Nutshell. In: Benferhat, S., Grant, J. (eds.) SUM 2011. LNCS (LNAI), vol. 6929, pp. 52–60. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23963-2_5
8. Vityaev, E.: Unified formalization of « natural » classification, « natural » concepts, and consciousness as integrated information by Giulio Tononi. In: The Sixth international conference on Biologically Inspired Cognitive Architectures, BICA 2015, Lyon, France, 6–8 November 2015, vol. 71, pp 169–177. Elsevier (2015). Procedia Computer Science
9. Kolonin, A.: Computable cognitive model based on social evidence and restricted by resources: Applications for personalized search and social media in multi-agent environments. In: International Conference on Biomedical Engineering and Computational Technologies (SIBIRCON) 2015, Novosibirsk, Russia (2015)
10. Kolonin A., Vityaev E., Orlov, Y.: Cognitive architecture of collective intelligence based on social evidence. In: Proceedings of 7th Annual International Conference on Biologically Inspired Cognitive Architectures BICA 2016, NY, USA, July 2016

11. Kolonin, A.: Architecture of Internet Agent with Social Awareness. In: 8th Annual International Conference on Biologically Inspired Cognitive Architectures BICA 2017, vol. 123, 2018, pp. 240–245 (2017). Procedia Computer Science
12. Kolonin, A.: Adaptive experiential learning for business intelligence agents. In: Cognitive Sciences, Genomics and Bioinformatics (CSGB) - Symposium Proceedings (2016)
13. Kolonin, A., Shamenkov, D., Muravev, A., Solovev, A.: Personal analytics for societies and businesses with Aigents online platform. In: 2017 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) - Conference Proceedings (2017)