

Sapient Systems in a Sandbox

Pablo Noriega

Abstract. The purpose of this Chapter is to set concrete grounds for two arguments. The first is in favor of approaching the notion of sapient systems from the unlikely perspective of stupidity. The second, in favor of paying serious attention to the environment where agents endowed with decision autonomy —be they sapient, human or otherwise— interact. Although the paper is ostensibly speculative, some concrete elements are advanced to substantiate both arguments.

1 Introduction

This is a position paper where I dare to propose a two-pronged strategy to approach sapient systems. The two research directions I advocate are summed up in the title of this paper and consist of working with individual sapient behavior, on one direction, and on the other, the collective, interaction environment. Both strategies have quite different concerns and conceptual realms, thus they may be appealing to quite different research communities. Each direction contributes its own scientific and technological background and brings, in particular, its own performance indicators to assess progress. These features are positive, I claim, because they may appeal to researchers with distant interests with the clear advantage that progress along either direction may be brought to bear upon the other.

To set the grounds for the topic I will advance, in the next section of this paper, a tentative characterization of sapient systems in terms of the space for innovation that internet and nanotechnology are opening for the long-held traditions of AI. The characterization involves three dimensions of sapient behavior that I will use to build my argument upon: self-sufficiency, adaptability and coordination. In Sec. 3, I pursue the argument for the study of a few features of individual intelligence that, I believe, are intrinsic to sapient behavior. I claim these features are inherently bounded and therefore a sensible approach to their artificial realization is to design simple-minded (stupid) sapient systems. In Sec. 4 I shift my attention to the intelligent behaviors involved in coordination and on the need for reifying a space where sapient action takes place. Taking advantage of what I did for individual features, I propose the components of an artificial environment that would provide a safe playing ground for low cardinality groups of stupid sapient agents.

2 Sapient systems in AI

2.1 Sapient systems

For the purpose of this paper I will adopt a simplified notion of a sapient system that shares the main features expressed in Mayorga (1999) which in my understanding are compatible with other characterizations of *sapient*, *smart* or *wise* systems (cf. Negrete-Martinez (2003); Mayorga (2005); Skolicki and Arciszewski (2005); Weigand (2005)). Thus:

Notion 1 *A sapient system is a computational entity that interacts with its environment, exhibits learning and adaptation behavior, is capable of creating knowledge and produces sound judgement.*

The vague terms I use in the characterization will become sharper as the discussion of the proposal proceeds but I will begin by pointing out some elements that shade my use of these terms. Note that in approaching the subject of smart systems Mayorga (for instance in Mayorga (1999)) draws from his background in AI, and softcomputing in particular, as well as the areas of control and operations research. My proposal, in turn, is biased by my own perspective from AI with a significant bias towards multi-agent systems and social psychology. In keeping with that bias, I am aware that my characterization sidesteps two aspects of sapiency that are fundamental in some of the other characterizations. Namely, the physical realization of a sapient system and the actual implementation of sapient behavior. I claim that neither are fundamental to the proposal which is the focus of this paper.

2.2 An AI turning point?

Classic AI is sometimes described as the discipline concerned with the computational simulation of *cognitive processes* and the creation of artifacts that are adequate stand-ins for intelligent behaviors. It is a convenient definition for it corresponds to the manifest programme of some of the founders of the field (in spite of the difference in emphasis and commitment that individual groups might have, McCorduck (1979)) and also with the outcomes of fifty years of activity in the field. For it may be argued that even if “general intelligence” has not been achieved, considerable progress has certainly been made in the understanding of specific cognitive processes like problem solving, natural language understanding or learning, and, significantly, in the engineering of artifacts that exhibit competent isolated cognitive behaviors. The success, we may note, has been in the plurality of those cognitive processes that have been synthesized.

Perhaps, though, we should also note that methodologically and conceptually the first fifty years of AI have focused, mostly, in the intelligent behavior of individuals. But as I will argue, that individualist focus is shifting towards a social view of intelligence. That change of focus has at least three significant drivers: internet, embodied and ubiquitous computing, and nanotechnology.

Internet brings forth the awareness of social intelligence. A digital reality in which transactions involve digital objects and are carried out in a digital –virtual– environment. But also a social reality in which interactions may involve participants who may

be numerous, geographically distant, ever-present and may be digital entities themselves. A reality, hence, that finds ideal tools in classic AI technologies but also a reality that gives AI splendid opportunities to innovate.

Embodied and ubiquitous computing is already proving to be a good example of those opportunities for AI. It shares with internet the relevant features of a digital-intensive environment with numerous participants. It involves significant inter-device activity but, as internet, it also is human-intensive in as much as the activity involves human users as the main beneficiaries. While we are already witnessing the first applications of ad-hoc networks, wireless communications and mobile devices, we can bet on a rapid adoption of cutting edge AI as well as conventional AI technologies that will expand the number and types of uses of embedded and pervasive computing.

Nanotechnology moves the boundary of artificial interactions to the nano-scale, thus opening the path not only of minute processors, but also for the design of devices that act *collectively* at a micro or nano-scale.

These three drivers are creating a wide innovation space for AI. The challenge is to build on top of the mature constructs of AI artificial systems that are able to coordinate with other systems (natural or artificial) to achieve a collective task, to be able to adapt and continue to operate under adverse circumstances and within dynamic environments while at the same time being able to rely on their own resourcefulness to achieve their tasks within a social milieu. Consequently, those three drivers bring into focus the need to study cognitive behavior as it happens within a social environment: the need to study not only individual behavior but *collective behavior* as well. In particular, instead of just focusing on *individual cognitive processes* as its primitive task, AI is poised to explore the wider notions of *interaction* and *environment*.

Ideally, sapient systems would be an appropriate artifact for the new AI because, ideally, sapient systems would be able to contend with the type of rational behavior needed to interact in a social environment. They should be adaptive, self-reliable, socially-aware and we would expect them to be competent in the tasks involved in collective action: negotiation, team making or conforming to an organization. The conceptual and engineering task of developing them may prove formidable, but perhaps a modest approach like the one I suggest in this paper might prove fruitful.

3 In praise of foolishness

Notion 1 was a first approximation to what I understand a sapient system might be. I may now add more detail. First I take advantage of the agent paradigm and assume a sapient system will be somewhat like an autonomous agent. Second, I avoid solipsism and will assume sapient interact with other sapient in social tasks that are as complex as the ones mentioned in the previous section. Hence, the three features of social interaction that I mentioned above –self-reliability, adaptability and coordination– should be part of the behavioral capabilities of a sapient, and I just paraphrase them in terms of likely operational realizations of each feature. Third I need to give some substance to the vague requirement of “creating knowledge and produce sound judgement”. For that I will invoke the notion of “insight” and make it extensionally clear by enumerating the

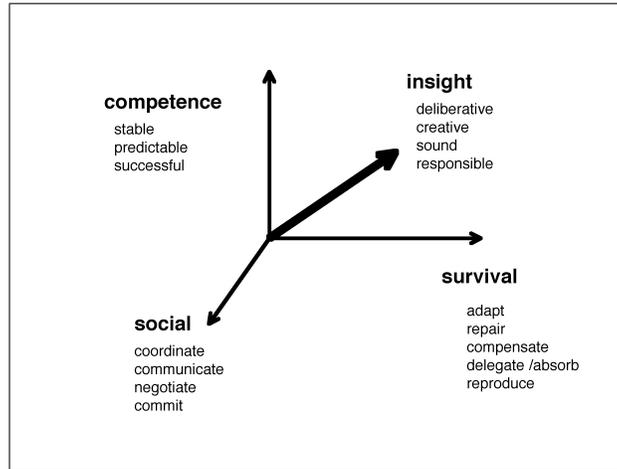


Fig. 1. A framework to describe the type of cognitive behaviors that a sapient agent should exhibit.

type of insightful behavior expected from a sapient system. The result is the following characterization that is represented in Fig. 1. Note that in this characterization I intend to profit as much as needed from classical AI developments and techniques by making explicit some capabilities that might be implemented in different ways.

Notion 2 A sapient system is an autonomous entity whose social behavior entails a combination of capabilities that may be classified along four dimensions: (i) Competence: stable, predictable, successful, ... (ii) Social: coordinate, communicate, negotiate, commit, ... (iii) Survival: adapt, repair, compensate, delegate and absorb, reproduce, ... (iv) Insight: deliberate, innovate, validate, assume responsibility, ...

I would like to make the notion of sapient system operational and for that purpose I will first propose a strategy by which we may approximate sapient proficiency in an “asymptotic” way.

Notion 3 A simple-minded agent is a computational implementation of a sapient system that has defective capabilities or lacks some of them or fails in putting them properly into practice.

In practice, stupid or simple-minded agents can be thought of as deliberative agents with distinct capabilities that are known to be defective to a certain degree. Capabilities that one would expect from a sapient system –like learning, planning, ontology matching or argumentation-based reasoning– may or may not be included by design in a given simple-minded agent and capabilities would be proficient up to a certain assessable degree. The intuition is that knowing in advance the composition and proficiency of primitive and compound capabilities would facilitate training and tailoring of simple-minded agents and be conducive to an overall behavior that is competent up

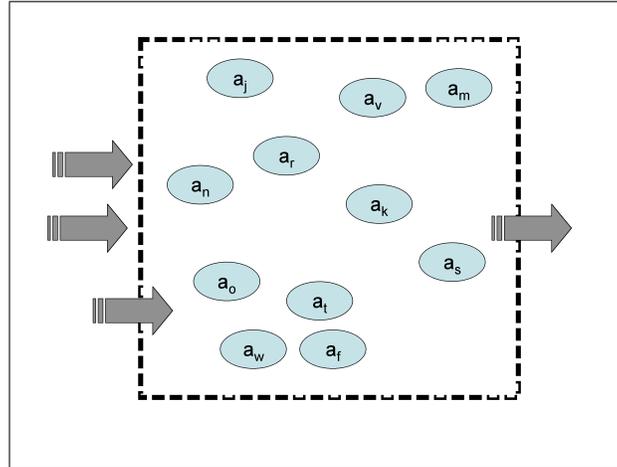


Fig. 2. A *simple-minded* sapient agent is a regulated MAS where conventional cognitive agents perform collective activities subject to explicit enforceable conventions.

to a certain objective degree. The same could be said about meta-capabilities –if we prefer to distinguish them from simpler more atomic capabilities for methodological or ontological reasons– like flexible goal attainment, achievement motivation, moral attitude, survival behavior. I should point out that dealing with meta-capabilities has two added advantages: it allows the design of control mechanisms that result or affect emergent behavior, like swarming or moral disposition, and also throws light on the possibility of task-dependent design of sapient agents by establishing competence measures –like the degree of docility of stupid agents towards other agents commands, or their stamina towards adverse outcomes– that are relevant for a given problem task.

I claim that simple-minded sapience, as just described, is a realistic project. Building deliberative agents with lacking or deficient cognitive capabilities is consistent with the tenant of bounded rationality, with the explicit advantage of setting the bounds at design time. It is conducive to incremental attainment of competence since intended degrees of accomplishment are made explicit at design time and validated at run-time. Ultimately, each capability would be stable and predictable. Moreover, building deliberative agents with lacking or defective meta-capabilities of the sort mentioned above is also advantageous from a methodological perspective since they facilitate addressing separately the tasks of designing tactic and strategical behaviors.

Since one can start building simple-minded agents from existing AI artifacts, this proposal would suggest an ideally monotonic process that approaches sapiency as the different behaviors are improved and more capabilities (and meta-capabilities) are modeled along the proposed four dimensions of sapient behavior.

The last element I need in order to give a characterization of a simple-minded sapient system is a device to control the actual behavior of the simple-minded agents. For that I rely on the notion of *regulated multi-agent system* (Fig. 2). A regulated MAS

is a collection of agents whose interactions comply with a set of conventions that “regulate” them. In a regulated MAS, there is some mechanism that makes participating agents held accountable for their actions. Depending on the type of mechanism and its implementation, the enforcement of conventions may be more or less strict, and unwanted outcomes restrained more or less effectively (cf. for example the COIN workshops proceedings Boissier et al. (2006) for a sample of alternatives).

In keeping with the idea of a monotonic approach to sapiency, I want to be able to impose a strict control over a simple-minded agent’s behavior; hence choose a regulatory mechanism that is strict and effective. Even more, I want one that may be designed in advance and then adapted –or adapt itself– to become better suited to guarantee the ostensible behavior. In this manner the regulated MAS would be the *liable* outcome of an aggregation of “intelligent” capabilities and meta-capabilities orchestrated by the regulatory environment. If a simple-minded agent is built as a regulated MAS whose conventions are properly enforced, the only ostensible behavior of the MAS is that which is admissible by the conventions. Thus allowing an incremental “damage control” on individual simple-minded agents by tuning the regulations to the expected outcomes and tuning agent capabilities to the desired levels of proficiency.

I may now make my proposal for sapient systems explicit:

Proposal 4 (P_{MAS}) *Sapient systems are feasible as multiagent systems:*

P_{MAS_1} : *Sapience is a complex quality that results from a collection of interrelated capabilities whose proficiency within some objective thresholds may be assessed.*

P_{MAS_2} : *Sapient systems may be approximated by building simple-minded sapient agents as software agents –proficient in specific capabilities or meta-capabilities– in a regulated multiagent system.*

To clarify further what I take sapient systems to be, I may now make explicit the other assumptions I am making.

Notion 5 *Assumptions about sapient systems:*

Boundary. *I assume a given sapient system is an individual distinguishable from other sapient systems and its environment. Hence it has some permanence, even if it may vary or evolve over time. I do not commit to a physical presence for in some cases a sapient may be a virtual entity that may be present, acting, in two different situations simultaneously.*

Mental. *Behavior is determined by abstract processes that correspond to cognitive processes of the type living systems have.*

Composition. *In what follows I will assume that a sapient system is a system whose ostensible behavior is an aggregation of capabilities along the four dimensions included in Notion 2.*

Achievement. *Behavior is expressed in observable actions and these actions may be successful or not in achieving objective goals. The degree of achievement may be measured directly or through objective indicators that subsume the effects of combined or complex behaviors.*

Artifact. In what follows I will assume that a sapient system is a computational system that may be situated in a, possibly, virtual environment and be able to interact with the environment which may include other sapient systems. That environment in some cases may also be engineered in advance as a regulated environment for autonomous agents (that will be the case for the sandboxes I discuss next).

Confinement. Ostensible behavior is mediated by a regulated interface –which is part of the sapient system, not of the environment– that determines the afferent and efferent information. That is, the regulated interface determines, in a strict way, what information from the environment may be received by the sapient system and what outcomes of the sapient system deliberations are eventually expressed into the environment.

Implementation I assume simple-minded agents are feasible in principle. However, I am not concerned by the way actual capabilities and meta-capabilities are implemented, nor what the actual architecture of the MAS is, nor how the confinement is achieved. Second, I am also avoiding any commitment as to the physical realization of sapient systems, I adduce that the abstract problem is interesting enough and a software implementation would provide rich enough models for the theory of sapient systems and useful enough artifacts to be used in sapient applications.

In consonance with what I suggested for capabilities, the design of the regulated environments may start from existing MAS technologies. In particular from the work on *electronic institutions* like the IIIA model and tools described in Arcos et al. (2005) and their extensions into autonomic ones as proposed in Bou et al. (2007).

4 Sapient in a sandbox

I will now turn to the second part of my proposal: environment engineering, or setting up a *sandbox* where simple-minded sapient agents are allowed to interact. I claim that this second element would make simple-minded sapient agents operational.

Proposal 6 (P_{env}) *Sapient systems are competent within a problem environment:*

P_{env_1} : *Sapient systems are situated in an dynamic environment whose state may be constantly changing but within a stable framework.*

P_{env_2} : *Sapient systems are purpose-explicit in the sense that their overall behavior is deemed competent with respect to that purpose and competence is measurable through explicit performance indicators.*

P_{env_3} : *A group of simple-minded agents may exhibit sapient behavior when interacting in a regulated structured multiagent environment.*

With the P_{MAS} proposal I argued that the notion of a stupid sapient agent is operational since it would involve modular deployment of simple minded components on the basis of available AI technologies and having explicit performance indicators would allow fine-tuning of the deployment. However, P_{MAS_2} introduced a design assumption that involves organizing many simple-minded components into an integrated MAS with the peculiarity of that MAS being *regulated*. What I am in fact assuming

is that stupid sapient behavior is the social behavior achieved by the interactions of the agents that constitute the MAS with the specific proviso that those interactions are subject to the explicit conventions that regulate the MAS.

In this section I elaborate further the schema of a regulated environment so that a collectivity of simple-minded agents is set loose in another –ideally less strict– *regulated structured environment* that captures the problem domain where sapients are intended to work and reflects the environmental conditions that would ideally prevail in the problem domain.

Notion 7 A Regulated Structured Environment *involves two main components;*

Object environment with a fixed ground ontology (agents, objects involved in actions, actions, ...), fixed pragmatics of atomic actions (conditions, effects) and a collection of social, interaction, regulatory and achievement structures that constrain or articulate interactions. Agents interact in that environment, all their actions are observable and all environmental constraints and regulations are enforceable

Meta-environment that allows the observation of object environment interactions, and measure the outcomes of those interactions. Furthermore, the meta-environment permits the sapient designer (or supervisor) to act upon entities of the environment and in particular modify the population of agents in the environment or change conventions and structures if needed.

Thus, the notion of a regulated structured environment (RSE) involves an *object environment* where stupid sapients act. They interact among themselves but also with other entities that are part of the environment, like stoplights and highways, tumors and organs, mobile phones and medical devices,... The environment is *structured* because, as part of the environment, there are coordination devices, regions, organizational structures, legal or normative resources that affect the way those interactions take place. For instance certain traffic regulations affect a sapient agent only when that agent is moving in a specific region, or a group of sapients may team-up to deal with a contingency invoking a particular type of contract that is available in the environment, or may choose to settle disagreements by voting according to specified protocols. This object environment involves elements and structures than were not available in the regulated MAS described above.

Regulated structured environments also have a *meta-environment* whose purpose is to observe the object environment and intervene on it. Thus it involves parameters that may be directly manipulated, behavioral and performance variables that may be observed, means to define observable indicators that synthesize outcomes of complex social interactions. In addition, the meta-environment includes different tools and mechanisms to set-up and register experiments and simulations, like deploying populations of sapient agents with specific features, automated generation of environmental conditions, ways of changing the structure of the environment or of its regulatory aspects –adding more objects, changing rules and conventions, for example.

Thus, it is in that sense that a RSE is a “sandbox” where stupid agents are left on their own contending with an artificial world whose creator has control over all that is part of it and may see everything that happens in it.

The agent-based simulation community has been developing agent environments that may serve as a basis for such sandboxes (cf. Weyns et al. (2007) for a variety of examples).

Having regulated structured environment to deploy stupid sapient in them would be rather useful since RSE could be use as simulation environments to experiment, test and tune sapient. Stupid sapient would become competent in the problem and might be then deployed and left to their own resources. Analogously, RSE may evolve and adapt to the problem domain at run-time (Bou et al. (2007)) and constitute a model of a stable problem environment where sapient systems are intended to act (e.g. a RES to manage traffic control policies in a large metropolitan area; or a RES with simple minded agents injected in a high-risk patient as a nano device to deal with contingent blood clots).

5 Closing remarks

I have presented a two-pronged strategy to address the design and development of sapient systems. On one dimension I suggested to build *simple-minded* agents whose cognitive capabilities and meta-capabilities are limited by design. Sapient behavior, I postulated, would result from the *collective interaction* of conventional agents that interact within a *regulated* social environment. The resulting regulated multagent system could be reified as a *stupid*sapient system and could in turn be immersed in a regulated environment so that many such sapient systems could interact within a confined environment –a *sand box*– whose features and properties are established at design time and may be enforced at run-time.

Sapient systems are asymptotically clever since each sapient MAS by changing the participating agents may involve as many and as varied cognitive capabilities as wanted, on one hand. On the other, the regulations that are imposed by the environment bind the activity of each simple-minded agent to admissible interactions, thus the collective outcomes are those that the conventions define and the conventions admit. Regulations themselves are object of design and consequently ostensible sapient behavior is tuned by tuning the simple-minded agents and the regulations. In particular, recent work on norm evolution and autonomic electronic institutions (e.g. Sierra et al. (2003); Bou et al. (2007, 2006)) indicate promising lines for future development.

I argue that stupid sapient, as I have proposed them, come close to the intuitive notion of sapient discussed in literature (cf. Mayorga (1999, 2005); Negrete-Martinez (2003)) with two convenient properties: (1) Functionality is asymptotically proficient and (2) Resulting sapient behavior is restriction compliant. In this way, simple-minded sapient would address the limitations of control and stability that Mayorga pointed out, and the social behavior of the MAS overcomes the AI-clumsiness outlined by Skolicki and Arciszewski (2005) while remaining independent from Weigand's DBI vs BDI propensities (Weigand (2005)).

I propose to embed simple-minded sapient in a (second order) regulated environment so that one can study their complex social interactions within a framework that facilitates observation, experimentation, and control of the social outcomes. The use

of the sandbox is twofold: it can be used to test and deploy sapient agents that will eventually be left on their own resources in an open environment, say in web service deployment. Or for some applications, like for medical nano-therapies, the useful artifact would be the sandbox together with its sapients.

The intuition beneath this proposal is that one can approach the ideal of a general intelligence, that has remained elusive for AI, through social processes. Or, put in other words, once we bring into the AI agenda the notion of collective cognitive behavior and social intelligence we are again facing the challenges that were at the root of AI. I believe, therefore, that achieving even limited sapience a worthy contemporary AI challenge.

Acknowledgments.

This work has been partly funded by project IEA (TIN2006-15662-C02-01). I would like to thank René Mayorga and José Negrete for their invitation to give a talk in the Workshop on Sapient Systems of the 2006 Mexican International Conference in AI (November 13, 2006). This paper is a revision of that conference and profits from their comments.

References

- Arcos, J. L., Esteva, M., Noriega, P., Rodríguez-Aguilar, J. A., and Sierra, C. (2005). Environment engineering for multiagent systems. *Engineering Applications of Artificial Intelligence*, 18 (2): 191-204.
- Boissier, O., Padget, J., Dignum, V., Lindemann, G., Matson, E., Ossowski, S., Sichman, J. S., and Vazquez-Salceda, J., editors (2006). *Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems AAMAS 2005 International Workshops on Agents, Norms and Institutions for Regulated Multi-Agent Systems, ANIREM 2005, and Organizations in Multi-Agent Systems, OOOOP 2005, Utrecht, The Netherlands, July 25-26, 2005, Revised Selected Papers*, volume 3913 of *Lecture notes in computer science*. Springer Verlag.
- Bou, E., Lopez-Sanchez, M., and Rodríguez-Aguilar, J. A. (2006). Norm adaptation of autonomic electronic institutions with multiple goals. *International Transactions on Systems Science and Applications*, 1(3): 227-238.
- Bou, E., Lopez-Sanchez, M., and Rodríguez-Aguilar, J. A. (2007). Towards self-configuration in autonomic electronic institutions. In Noriega, P., Vazquez-Salceda, J., Boella, G., Boissier, O., Dignum, V., Formara, N., and Matson, E., editors, *Coordination, Organizations, Institutions and Norms in Multi-Agent Systems II. Revised selected papers from the COIN workshops held in AAMAS 2006 (Hakodate, Japan) and in ECAI 2006, (Riva del Garda, Italy)*, volume 4386 of *Lecture Notes in Computer Science*, In press. Springer Verlag.
- Mayorga, R. V. (1999). Towards computational wisdom: Intelligent/wise systems, paradigms, and metabots. In *Tutorial Notes. ANIROB, Congreso Nacional de Robotica, CONAR99*, Cd. Juarez, Mx.

- Mayorga, R. V. (2005). Towards computational sapience (wisdom): A paradigm for sapient (wise) systems. In Mayorga, R. V. and Perlovsky, L. I., editors, *2005 IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems KIMAS'05*, pages 4–12, Boston MA, USA. IEEE Press.
- McCorduck, P. (1979). *Machines who Think*. W.H. Freeman and Company, Washington, DC.
- Negrete-Martinez, J. (2003). Paradigms behind a discussion on artificial intelligent/smart systems. In *International Conference on Integration of Knowledge Intensive Multi-Agent Systems. KIMAS 03: Modeling, Exploration and Engineering*, pages 392 – 394, Boston MA, USA. IEEE Press.
- Sierra, C., Sabater-Mir, J., Agusti-Cullell, J., and Garca, P. (2003). Integrating evolutionary computing and the sadde methodology. In Rosenschein, J. S., Sandholm, T., Wooldridge, M., and Yokoo, M., editors, *Second International Conference on Autonomous Agents and Multiagent systems (AAMAS-03) July-2003 Melbourne , Australia*, pages 1116–1117. ACM Press.
- Skolicki, Z. and Arciszewski, T. (2005). Sapient agents seven approaches. In Mayorga, R. V. and Perlovsky, L. I., editors, *2005 IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems KIMAS'05*, pages 48–49, Boston MA, USA. IEEE Press.
- Weigand, K. A. (2005). Toward wisdom in procedural reasoning: Dbi, not bdi. In Mayorga, R. V. and Perlovsky, L. I., editors, *2005 IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems KIMAS'05*, pages 56–62, Boston MA, USA. IEEE Press.
- Weyns, D., Parunak, H., and Michel, F., editors (2007). *Environments for Multiagent Systems III*, volume 4389 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag.