The Rational and Irrational Minds of Others

Sergei Nirenburg
Department of Cognitive Science
Rensselaer Polytechnic Institute
nirens@rpi.edu

1 Building Integrated Systems

When one surveys Professor Carbonell’s many and varied research contributions, two more or less constant methodological trends seem to stand out. They are the preference for an integrative outlook and insistence on developing computational systems. The integrative outlook stresses the multiplicity of causes in behavior and contributing factors in reasoning. It is a rather difficult position to maintain. Indeed, with ever-accelerating specialization in science, the temptation is to concentrate on a narrow research area where one has more chances to attain deeper understanding. Still, while pursuing research in planning, reasoning, language understanding or learning, Jaime, it seems to me, has maintained his original interest in developing comprehensive models of human cognitive capacities. This choice can be traced back to POLITICS (Carbonell, 1978), which addressed reasoning, memory, intentional behavior, language processing and learning. At the same time, Jaime has insisted on developing proof of concept systems and applications that made the underlying theories and models tangible. Notably, whenever possible he opted for building systems that carried explanatory power in addition to producing expected deliverables. Moreover, in the best traditions of “good old-fashioned AI” he preferred, again, whenever possible, to develop applications that aimed not only to attain ever more spectacular functionalities but to do so while modeling cognitive capacities of humans.

I have singled out the trends of integration-orientedness and system-building by applying analogical reasoning; these trends can be discerned in the research my team and I have been carrying out over the years as well. Quite possibly, this is due to a historical bias: after all, I worked in Jaime’s group for seven years in the 1980s and 1990s, notably, on the KBMT-89 knowledge-based machine translation project (Nirenburg et al., 1992). In what follows I will leave the etiology of this analogy aside and discuss some selected preferences and topics of our research for which POLITICS can be considered an inspiration.

Outside the work on generic architectures, the prevalent trend in the field is to separate the overall task of modeling cognitive processes into autonomous constituent subfields. This tactical decision is understandable because addressing the overall problem may require more resources of all kinds than are available. Indeed, the typical explanation of this choice is “this is how science is done.” One of the celebrated examples is the separation of syntax from semantics and semantics from pragmatics in theoretical linguistics, a choice that trickled down into computational linguistic work. The drawbacks of this compartmentalization approach certainly do not remain undetected: for example, it has been observed long ago that some syntactic dependencies (such as prepositional phrase attachments) can be fully resolved only when semantic knowledge is evoked. The perception of the impossibility of attaining a 100% analysis within a single processing module (e.g., syntactic parser) has added motivation for adopting probabilistic and statistical approaches to the task. We believe that irrespective of whether a processing algorithm is rule-based or statistical or a mixture of these approaches, it is beneficial to view the modeling of cognitive processes as an integral whole. Indeed, to give just one example, the syntax vs. semantics vs. pragmatics split in theoretical linguistics leaves phenomena like ellipsis only skeletally described, with select aspects being treated in isolation by different theories (see, e.g., McShane et al., 2012a for a detailed analysis).
The related issue of simplifications and assumptions - both explicit and implicit - in service of scientific and technological progress has also been at the center of our thinking. All programs of study involve simplifications and assumptions, but we believe that their nature and practical implications should be addressed more often and more rigorously than has been done in practice. Consider the following observations about recent work in NLP:

a. NLP based on supervised learning - which accounts for much of NLP over the past 20 years or so - covers only a relatively small fraction of language phenomena, that for which corpora have been annotated (for an overview of annotation efforts in NLP see, e.g., Palmer and Xue, 2010); moreover, far from all language phenomena lend themselves to effective and efficient corpus annotation;

b. the cost of annotating corpora, not to mention the cost of formulating the inevitably complex annotation schemata, is rarely discussed in the same breath as the purportedly low cost of statistical methods;

c. the fact that many processing engines are not fully automatic, since they rely on manually provided preconditions, is often not explicitly discussed, nor are the implications of this state of affairs discussed;

d. the assumption that single-task engines will ultimately be combined into larger environments has not been amply corroborated; and

e. the long-term prospects of many approaches are questionable: e.g., if a research paradigm is devoted to functioning in the absence of semantic heuristics, what becomes of it as more and more semantic heuristics are made available?

Inferring the intentions of agents and choosing a course of action on the basis of goal-directed inferencing was a major capability of POLITICS that separated it from earlier agent simulators, such as ELIZA (Weizenbaum, 1966) and belief modeling systems such as the Goldwater machine (Abelson, 1973). At the risk of oversimplification, one can draw an evolutionary path here from a pure stimulus-response behavior of ELIZA to the perception-reasoning-action sequence in POLITICS, where the methods of reasoning included a mixture of inductive, deductive and abductive inference rules together with at least a trace of reasoning by analogy. The latter is evident, for example, in the centrality of scripts as the core knowledge substrate for the system: agents interpreted actions (including verbal actions) of others and chose their own actions largely by matching input content representations against component actions in stored scripts.

While the mechanism of encoding knowledge was neutral in this respect, it was assumed that the sequence of events in scripts and plans and the agent’s inventory of goals and inference and decision rules must support rational reasoning. Hence, agents were tacitly expected to be at all times pursuing one or more of their goals and to prefer the most efficient of the available plans to attain these goals. Of course, an important prerequisite for goal-directed reasoning is supplying the agents with the ability to choose the most appropriate goals and plans to pursue in specific circumstances.

2 The Integrative Stance in OntoAgent

Some facets of our team’s work on developing language-endowed intelligent agents can be seen as an extension of this line of research. In this paper I will briefly introduce two of them – including beliefs about other agents in the knowledge substrate of intelligent agents and modeling irrational alongside rational behavior.

A core direction of work in our approach to agent modeling, the OntoAgent project (Nirenburg et al., 2010), is endowing agents with an increasingly encompassing and sophisticated computational theory of mind. Using this theory we model the OntoAgents’ insights into the minds of others as well as into their own mind. The theory of mind underlying OntoAgent assumes that the knowledge the various OntoAgents possess is not identical but has a non-trivial intersection, with agents capable of honing their models of others through observation, reasoning and interaction. An agent’s ability to reason about itself is known as metacognitive ability. An agent’s ability to use models of the minds of others has come to be referred to as “mindreading” (admittedly, a rather grandiose term).

The OntoAgent approach allows one to configure agents of different levels of sophistication and featuring different inventories of functionalities, metacognition and mindreading among them. For example, agents that are prepared for the role of virtual patient in our medical training system,
MVP (McShane et al., 2009), are endowed with a specific personality profile and a model of its current physical and mental state; a varying level of language-processing, decision-making, learning and mindreading capabilities; as well as a physiological model and the capability to perceive physiological symptoms (interoception). They may also be optionally endowed with some metacognitive abilities. By contrast, the agent that plays the role of automatic tutor in MVP is not supplied with a personality profile or a physiological model. In the current version of the system the tutor is not endowed with metacognitive or mindreading capabilities either.1

Metacognitive and mindreading processes in OntoAgent take place after other processing modules of an intelligent agent a) produce the semantic representation of an input, such as a dialog turn; and b) augment this representation through processing a variety of implied meanings not directly observable in the inputs. At this point, in addition to updating the record of the current situation, Agent (A) may choose to “mindread” the producer (P) of the input through:

a. Belief ascription: modifying A’s model of P’s world model as a result of processing the input (“P thinks that the Earth is flat” or “P does not trust me”);

b. Goal ascription: modifying A’s model of P’s goal agenda or inventory (“P wants to learn to juggle”); note that this understanding affects decisions about what implicatures to derive and when, how to resolve vagueness and ambiguity in inputs, etc.;

c. Character trait ascription: inferring and adjusting the values of A’s model of P’s character traits (stoicism, arrogance, etc.), personal preferences and quirks (“P is tridecaphobic”); and

d. Detection of potential decision-making biases (see below).

Mindreading results in better understanding of other agents. This is beneficial for two reasons. First, results of (successful) mindreading supply the mindreading-capable OntoAgent with valuable additional decision-making knowledge. For example, an advice-giving agent may choose different advice when addressing a preschooler than when talking with an adult. Moreover, the same advice will be formulated differently depending on what the agent believes that the advisee already knows about the topic of advice or the advisee’s reading ability. Second, results of (successful) prior mindreading help the interpretation of subsequent inputs from the same agent.

3 Modeling Humans, Not Superhumans

Here is a brief illustration. In one of our experiments with MVP, we created several OntoAgents with different theories of the mind of their interlocutor, the (human) physician. When, during a diagnostic interview, the physician asked: “Have you been traveling recently?” those OntoAgents that a) were not medically sophisticated; and b) believed that the physician had a good personal rapport with them tended to respond directly without taking into account the diagnostic situation, e.g., “I drove down to see my daughter in Philly last weekend.” Those agents that were medically sophisticated, irrespective of their belief about the physician’s rapport with them, responded “to the point,” e.g., “I have not been in any place where the outbreaks happened.”

Since we are modeling human behavior and not developing superefficient, superintelligent agents, we must also address situations resulting from unsuccessful, incorrect mindreading. The response from the medically unsophisticated agent above is an example of an incorrect mindreading instance. Having received this input from the agent, the physician should recognize its inconsistency with the goal he or she was pursuing in asking the original question. This may lead the physician to realize that the agent carried out an unsuccessful instance of mindreading. (Inconsistencies of this and other kinds are useful triggers not only for detecting incorrect interpretation of inputs but also as clues for further mindreading. For example, an inconsistency between a patient’s report and his or her test results may lead the physician to conclude that the patient is lying – a typical eventuality in actual doctor-patient interactions. See McShane et al., 2012b for a detailed discussion of inconsistencies as diagnostics in OntoAgent.)

But what if, for whatever reason, the physician fails to recognize faulty mindreading on the part of the patient? This may unnecessarily delay the diagnostic process. So, it would be useful if the physician got help in recognizing the situation of

1 In addition to the components listed, the knowledge substrate of OntoAgents includes, as should be expected, a model of the world, a language model, a model of the situation at hand and a memory of specific events.
faulty mindreading. We are working on an application of OntoAgent called CLAD, a CLinician's Advisor, that observes an interaction between a physician and a patient and gives support and advice to the physician, working essentially as the latter’s cognitive prosthesis. Note that CLAD must maintain more complex models of agents’ minds than MVP. MVP, like most systems that model belief ascription (e.g., ViewGen, Ballim and Wilks, 1991) deal with single-link belief ascription (A’s beliefs about B’s beliefs). CLAD represents and maintains models of the minds of the patient and of the physician. Due to the nature of CLAD’s task, these models must also include CLAD’s model of the physician’s model of the patient’s mind and the patient’s model of the physician’s mind.

In developing CLAD, we are also studying human behavior that cannot be called rational. Specifically, we are working on detecting agent biases. Biased thinking—such as jumping to conclusions or letting fatigue affect one’s judgment—is a pervasive feature of human decision-making (Kahneman, 2011). Our hypothesis is that bias-influenced judgment errors could be reduced if CLAD were able to detect potentially biased decisions and generate explanatory alerts to the physician. Among the biases we currently address are depletion effect, jumping to conclusions, overconfidence, false intuitions, base-rate neglect, the illusion of validity, the halo effect, the framing sway and some others. (For a detailed discussion of our approach to detecting biases see McShane et al., 2013.) Here I will present just one brief illustration.

Consider the base-rate neglect bias. It is a type of decision-making bias that, applied to clinical medicine, can refer to losing sight of the expected probability of a disease for a given type of patient in a given circumstance. For example, a patient presenting with a fever to an emergency room in New York is highly unlikely to have malaria, whereas that diagnosis would be very common in sub-Saharan Africa. Although physicians are trained to think about the relative likelihood of different diagnoses, remembering all of the relative probabilities can be quite challenging, given different constellations of signs and symptoms and in the conditions of time pressure, stress and fatigue (that cause depletion effect biases). CLAD can help with this by detecting situations in which a doctor is pursuing a diagnostic hypothesis that is unlikely given the available data.

For example, esophageal carcinoma can result from gastroesophageal reflux disease (GERD) but typically only if GERD has not been successfully treated for over five years and if the person checks for at least one of the following features: age over 40, smoking, regular alcohol intake, residence near or employment in an industrial environment or exposure to carcinogenic materials. These likelihood conditions are recorded in the ontology as a composite filler for the property SUFFICIENT-GROUNDS-TO-SUSPECT for the disease ESOPHAGEAL-CARCINOMA. If a physician hypothesizes esophageal carcinoma for a 20-year old nonsmoker with a 3-month history of GERD, CLAD detects an instance of base-rate neglect bias and will issue a warning that there appears to be insufficient evidence for this hypothesis, and it will show the clinician the conditions under which the hypothesis is typically justified.

Work on OntoAgent seeks to integrate a broad variety of topics, tasks and “microtheories.” The purpose of this brief note is to illustrate how the philosophy and methodology behind some of Professor Carbonell’s early work is still very much on agenda in the minds of others.

References


M. McShane, S. Nirenburg, and B. Jarrell. 2013. Modeling decision-making biases. Biologically-


