AAAI Fall Symposium on Knowledge and Action at Social and Organizational Levels, 1991

Research Patronage and Distrusting A.I. Agents: Similar Problems with Similar Solutions?

by Robin Hanson
Artificial Intelligence Research Branch
NASA Ames Research Center, MS-244-17
Moffett Field, CA 94035, USA
hanson@charon.arc.nasa.gov 415-604-3361

The research patron's problem is how a non-expert can spend a given amount of funding so as to induce the most scientific progress toward answering questions of interest to that non-expert patron. Along the way, such patrons often want to induce consensus estimates about where the weight of evidence lies on such questions, to guide the patron's external actions. Multiple autonomous A.I. agents, who distrust each other, also want trustworthy ways to induce one another to create and share information. And, if possible, they want to form consensus estimates they all act as if they agree with, to avoid inconsistency costs. Are these problems related?

THE PATRON'S PROBLEM

History and philosophy of science, both normative and descriptive, have traditionally focused on cognitive issues, neglecting the social dimension of science. The recent surge in social studies of science has corrected this somewhat, except that social studies have shied away from dealing with the normative issues of how science should be organized.

As a result, science patrons, both in government and private foundations, are left somewhat adrift amid a storm of self-interested advice. What relevant literature there is focuses largely on small variations of existing institutions, such as in citation analyses, peer-review agreement statistics, econometric studies, small experiments, and over-simplified formal economic analyses.

Yet, contrary to popular impression, there is little reason to think that our current academic institutions are near optimum. Scientific funding and consensus methods have varied dramatically over the centuries and across the world, including direct patron/scientist relations, prizes, large hierarchical civil services, scientist self-patronage, self-evaluation within professional societies, etc.

Public patrons like the U.S. Congress are taking an increasingly skeptical eye toward existing institutions, and the time is ripe for a more fundamental rethinking of our methods of funding and consensus. In fact a number of interesting radical proposals have recently been made. A.I. hubris and engineering-orientation may be just what is needed to break out of traditional patterns and consider the patron's problem anew.

DISTRUSTING A.I. AGENTS

A powerful approach to building intelligent systems is that of modularity and subsumption -- build "societies of mind" with bigger smarter agents composed of smaller stupider agents. Having each agent look out for itself may make the system more robust in the face of individual agent failures and irrationalities, allowing more incremental and experimental improvement.

But agent autonomy introduces various costs of skepticism, since agents must now negotiate with each other and cannot naively trust what other agents tell them. Reasoning which crosses agent boundaries needs to be mediated by institutions which discourage lying, and the system as a whole will pay inconsistency costs unless each component agent can be induced to act as if they agree with some total consensus, even if they privately disagree with it.

Formal analyses of dramatically oversimplified problems, like blocks-world or worse, can mislead us about what the important problems are and what are scaleable solutions. Practical solutions to the patron's problem, however, allowing real distrusting humans to engage in knowledge production in a knowledge rich context, should teach A.I. folks a thing or two.

A COMMON SOLUTION?

An intriguing solution to both of these sets of problems is to simply introduce the core mechanism from a standard method of financial analysis: markets in contingent assets (really betting markets).

By introducing and subsidizing markets on particular science questions, a patron can both efficiently induce research on those questions and induce continuous and precise consensus estimates on them. The non-expert patron need make no choices about who should research a question or what methods are appropriate.

Distrusting A.I. agents can also use such markets to form a consensus on important questions, so that it is in each agent's interest to act as if they agree with the consensus. Such markets can also be used as trustworthy information channels; it is very hard to win by lying to the market (betting against one's beliefs) and if others have concerns, they can use the market to take out insurance, so the question becomes irrelevant to them.

This proposal is radical, but it deserves a closer look.

THE PATRON'S PROBLEM

[Cic] Cicchetti, D. "The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation", *Behavioral and Brain Sciences*, 14:1, pp.119-135.

[Ch] Chubin, D., Hackett, E., (1990) *Peerless Science*, SUNY Press, NY.

[Co] Cole, S., Cole, J., (1987) "Testing the Ortega Hypothesis: Milestone or Millstone?", *Scientometrics*, 12:5-6 pp.345-353.

[Gh] Ghiselin, Michael T. (1989) *Intellectual Compromise: The Bottom Line*, Paragon House Publ., NY.

[Red] Redner, H. (1987) "Pathologies of Science", *Social Epistemology*, 1:3, pp.215-247.

[ShS] Shapin, S., Schaffer, S. (1985) *Leviathan and the Air-Pump*, Princeton U. Press.

[SmP] Smith, P. (1990) *Killing the Spirit*, Viking/Penguin, NY.

[Syk] Sykes, C. (1988) *ProfScam*, St. Martin's Press, NY.

[Tul] Tullock, G. (1966) The Organization of Inquiry, Duke Univ. Press, London.

[Tur] Turner, S. (1990) "Forms of Patronage", *Theories of Science in Society*, ed. S. Cozzens, T. Gieryn, Indiana U. Press, pp.185-21.

RADICAL PROPOSALS

[By] Byrne, G. (1989) "A Modest Proposal", *Science*, 244, April 2, p.290.

[Gi] Gilmore, J.B. (1991) "On Forecasting validity and finessing reliability", *Behavioral and Brain Sciences*, 14:1, pp.148-149.

[Kan] Kantrowitz, A. (1977) "The Science Court Experiment: Criticisms and Responses", *Bulletin of the Atomic Scientists*, April, pp.44-50.

[Roy] Roy, R. (1985) "Funding Science: The *Real* Defects of Peer Review and An Alternative To It", *Science, Technology and Human Values*, 10:3 Summer, pp.73-81.

[Wa] Wade, N. (1980) "Why Government Should Not Fund Science", *Science*, 210:3, October, p.33.

CONSENSUS

[Man] Mann, C. (1990) "Meta-Analysis in the Breech" *Science* 249, August 3, pp.476-480.

[Gen] Genest, C., Zidek, J. (1986) "Combining Probability Distributions: A Critique and Annotated Bibliography", *Statistical Science* 1:1, pp.114-148.

[Gr] Grofman, B., Owen, G. eds. (1986) Information Pooling and Group Decision Making: Proc of the Second Univ. Cal. Irvine Conf. on Political Economy, JAI Press, Inc. London.

[Se] Seidenfeld, T. (1990) "Two Perspectives on Consensus for (Bayesian) Inference and Decisions" *Knowledge Representation and Defeasible Reasoning*, H. Kyburg, et. al. eds. pp267-286.

[Syn] (1985) *Synthese*, "Consensus" issue, 62:1, Jan

DECEPTION

[Cia] Cialdini, R. (1988) *Influence, Science and Practice*, Scott, Foresman and Co., Boston.

[Kah] Kahneman, D., Tversky, A., eds., (1982) *Judgment under uncertainty: Heuristics and biases*, Cambridge Univ. Press, NY.

[My] Myers, D. (1983) *Social Psychology*, 2nd ed., McGraw-Hill.

DISTRUSTING A.I. AGENTS

[Han] Hanson, R. "Even Adversarial Agents Should Appear to Agree", available from author.

[Ma] Malone, T., Fikes, R., Howard, M. (1988) "Enterprise: A market-like task scheduler for distributed computing environments", in B. Huberman, ed., *The Ecology of Computation*, pp.177--205. North Holland Publ. Co., Amsterdam.

[Mi] Miller, M., Drexler, E. (1988) "Markets and computation: Agoric open systems" in B. Huberman, ed., *The Ecology of Computation*, North Holland Publ. Co., Amsterdam.

[Ro] Rosenschein, J., Genesereth, M. (1985) "Deals among rational agents", Proceedings of the Ninth International Joint Conference on Artificial Intelligence, pp.91--99.

BETTING MARKETS

[Fo] Forsythe, R., Nelson, F., Neumann, G., Wright, J. (1990) "The Explanation and Prediction of Presidential Elections: A Market Alternative to Polls" *Economics Working Paper* 90-11, April 12. Univ. of Iowa, Iowa City.

[Han] Hanson, R. (1990) "Could Gambling Save Science? Encouraging an Honest Consensus" *Proc. Eighth Intl. Conf. on Risk and Gambling*, July, London.

[Hir] Hirshleifer, J. (1971) "The Private and Social Value of Information and the Reward to Inventive Activity", *American Economics Review*, 61:4, Sept., pp. 561-74.

[Kad] Kadane, J., Winkler, R. (1988) "Separating Probability Elicitation from Utilities" *J. American Stat. Assoc.*, June, 83:402, Theory and Methods, pp. 357-363.

[La] Laffont, J.J. (1989) *The Economics of Uncertainty and Information*, MIT Press.

[Na] Nau, R., McCardle, K., "Arbitrage, Rationality, and Equilibrium", *Theory and Decision*, to appear.

[ShW] Sharpe, W. (1985) *Investments*, 3rd Ed., Prentice Hall, NJ.

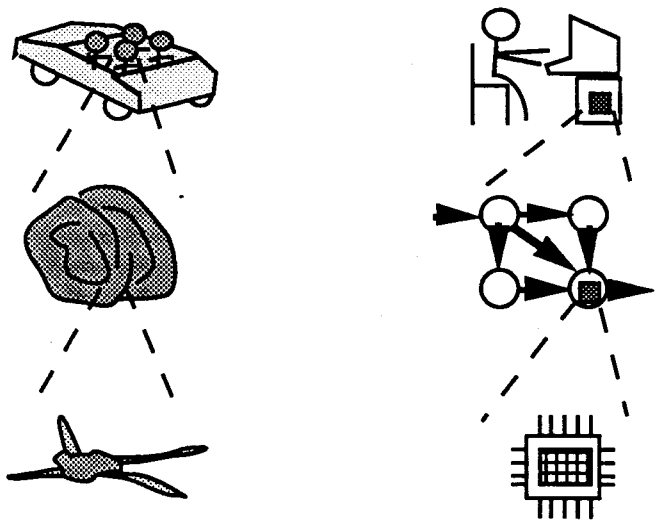
[Ze] Zeckhauser, R., Viscusi, W. (1990) "Risk Within Reason", *Science*, 248, May 4, pp.559-564.

QUESTIONS

These questions become increasingly speculative.

1) How can a non-expert patron of research spend a given amount of funding so as to induce the most scientific progress toward answering questions of interest to that patron?
2) How can such patrons induce sound consensus estimates along the way about where the weight of evidence lies on such questions, to guide the patron's external actions?
3) How can multiple autonomous A.I. agents, who distrust each other, find trustworthy ways to induce one another to create and share information, allowing inference to cross agent boundaries?

4) How can they, to avoid inconsistency costs, form consensus estimates they all act as if they agree with, even when they are autonomous and really disagree?
5) Consider composing intelligent systems from intelligent components, as in the figure below. At what level of aggregation do we want, or can we expect, systems to have consistent beliefs? Consistent goals? To act "rationally"? To be skeptical of each other? To compete?



6) What properties of systems do we want to hold at a wide variety of aggregation scales? If some properties are only appropriate at certain levels of aggregation, what makes those levels special?
7) To what extent does the availability of other co-existing systems allow us to weaken traditional notions of agent "rationality"?
8) How does the fact that real A.I.s, when we get them, will be imbedded in and can draw support from the same rich supporting culture, society, and economy that we are in change our conceptions about what will be hard and easy about designing such systems? That is, why should an A.I. need to know how to do something it can hire a human to do cheaper?
9) In a combined society of A.I.s and humans, what will humans have to contribute? How will this change as A.I.s get smarter?
10) What features of human society and modes of social interaction stem largely from peculiarities of humans technical limitations, and from human biological heritage, and should not particularly constrain societies of A.I.s (and perhaps modified humans)? For example, what if A.I.s need not "die" often, can easily copy themselves, can share and exchange parts of themselves, may have a much larger bandwidth for interaction, may be much more specialized for specific cognitive tasks, and need not be constrained by genetically programmed goals.
11) Can A.I.s use a more precise language with each other than humans do? Can they use formal contracts and markets more? Will they be less susceptible to national and cultural allegiances? Will they be more "cutthroat" in modifying their basic goals to whatever allows them to reproduce and/or be economically prosperous?
12) At different levels of aggregation, what will be typical degrees of commitment of systems to specific patterns of interaction with other systems? That is, will "agents" typically have little freedom because they were born with or contracted to work in a particular role within a particular highly structured organization, or will they usually be more autonomous, using credible threats to change agent relations as negotiating levers to induce competition?
13) Will "slavery" relations between A.I.s be typical? With so many possible agent "sizes", will economic measures replace head-counts in determining representation in political organizations?
14) What knowledge sources and tasks will typically be shared, with most agents contracting to gain access to this knowledge or reasoning as needed from standard sources through standardized interfaces, and what knowledge will typically be agent specific,created anew by each agent as needed, and perhaps so situation specific that its not clear it could be shared?