Intrinsic and Extrinsic Motivation in Intelligent Systems

Henry Lieberman LIEBER@MEDIA.MIT.EDU *MIT Computer Science and AI Lab*

Editors: Minsky, H. and Robertson, P. and Georgeon, O. L. and Minsky, M. and Shaoul, C.

Abstract

There are two ways that systems, human or machine, can get ”motivated” to take action in problem solving. One, they can be given goals by some external entity. In some instances, they might have no capability other than to work towards the goals provided by that entity. Two, they can have their own, internal goals, and work towards those goals. If given a goal by an outside entity, they can then try to figure out whether, and how, the external goal might align with their internal goals. In that case, the agent might be said to be acting in a ”self-supervised” manner. There are, of course, cases where both intrinsic and extrinsic motivation come into play.

This paper will argue that many machine learning systems, as well as human organizations, put too much emphasis on extrinsic motivation, and have not fully taken advantage of the potential of intrinsic motivation. Reinforcement learning systems, for example, have a ”reward signal” that is the sole extrinsic motivating factor. It is no wonder then, that even when such systems work well, they are incapable of explaining themselves, because they cannot express an explanation in terms of their own (or their users’) goals. In human organizations, relying only on extrinsic motivation (= ”incentive”) leads to rigid or dictatorial organizations; engaging internal motivation (at some cost to ”organizational efficiency”) can lead to creativity and invention.

Keywords: Intrinsic motivation, extrinsic motivation, goals, plans, actions, self-supervised learning

1. Motivation in people and machines

Nothing happens unless somebody is motivated to make it happen. If we want to understand events, we have to understand the motivation of participants in those events. If we want to make something happen, we have to make sure that the people involved are motivated to do it. So understanding motivation is essential for understanding human activity.

Artificial intelligence systems, also, need to have motivation. That’s not just a fantasy of anthropomorphization. We build artificial intelligence systems in our own image; so that many of the concepts that are applicable to human activity, are also applicable to activity by machines. We ask computers to help us in our own problem-solving activities, so the computer needs to understand what these activities are and why they take place, in order to be of help to us.

The cartoon caricature of a computer is that it just ”does what it’s told” by the programmer or user. It has no will of its own, and therefore no motivation. Of course, that’s true at some level, because the computer is just following program instructions.

But the situation is not that simple. As programs get complex, it’s impossible for programmers to foresee every detail of what the computer should do in various circumstances.

We would like to be able to specify our goals for the program at a high level, where we don't need to specify every detail. Computers are useful precisely because they can pay attention to myriad small details that we may not want to bother with. They are useful precisely because they may have access to knowledge or processing power that we may not have ourselves. They are useful precisely because we can delegate our goals to them, and they can work on achieving them, while we are free to do something else.

Of course, we don't want an AI to completely have a mind of its own, either. We don't want it to run amok and do things that are contrary to our interests. So we are always in an intermediate point between complete control and complete autonomy.

The best we can do is to tell the computer what it is we want, and program the computer with criteria and policies for making decisions. Then, when it reaches a decision point, or novel situation, it doesn't have to go back to ask a human to figure out what to do next. That is, we have to give it motivation. But what kind of motivation?

2. Why Can't We All Just Get Along?

Please forgive me a brief digression. In recent years, I have been becoming more and more concerned with big questions of society, and the role of AI in it. Why do we have war, poverty, climate change, and other ills? Is it just due to greed and stupidity on the part of humanity? Will we ever solve these problems? Will AI be part of the problem or part of the solution?

I got into AI because I believe that AI can be a force for good – that it can make major contributions to solving humanity's biggest problems. Rather than assign blame to human personality failings, I think one of the fundamental problems is that we just don't know enough right now about how human beings work. We don't understand enough about psychology and human intelligence to be able to "fix the bugs" that cause major man-made disasters. That's actually good news, because science, and particularly AI, can help. As we learn more about human intelligence, and as we learn more about how to make computers help us produce the resources to enable everyone to have a good life, we can make progress. So, I'm very optimistic.

Together with my co-author, Christopher Fry, we've written a book [Fry and Lieberman (2018)], entitled "Why Can't We All Just Get Along". The theme of the book is cooperation versus competition. We claim that most of the world's problems stem from people competing when they should be cooperating. We give arguments from game theory, based on the Prisoner's Dilemma [Axelrod (1984)], and from positive-sum evolution [Wright (1994)]. We argue that the impetus for competition is driven by scarcity of resources. Which, again, is good news. Because AI and other technologies like 3D printing hold the promise of solving scarcity, thereby enabling a more cooperative society. These are bold claims, and way beyond the scope of this paper, but I encourage you to read the book (or see the TEDx talk) at http://www.whycantwe.org/.

This quest has led me more and more to consider issues of human motivation. Why are people motivated to do things that turn out to be good, or turn out to be bad? What causes antisocial behavior, and what are the motivations of criminals? Why do people seek unlimited power and money even beyond what is beneficial for them? What is the motivation for needless and destructive competition?

Why Can't We All Just Get Along?
HOW SCIENCE CAN ENABLE A MORE COOPERATIVE FUTURE
Christopher Fry
Henry Lieberman

Figure 1: Why can't we all just get along?

3. Intrinsic vs. Extrinsic Motivation

Key to understanding how motivation affects competition and cooperation, is the distinction between intrinsic and extrinsic motivation.

Intrinsic motivation is like listening to music. You do it because you enjoy it for its own sake. Nobody has to force you or pay you to listen to music. Extrinsic motivation is when you are motivated for reasons other than the activity itself – you are paid a salary to work at a job, you get a reward, a gold star, social status, etc. (of course, some situations have both kinds of motivation).

Our present society is overwhelmingly imbalanced towards competition rather than cooperation – in our Capitalist economy between companies and products, and in our democratic government between candidates and parties. As a result, society is organized around extrinsic motivation. Government is about extrinsic motivation via laws and regulations; the economy provides extrinsic motivation in the form of financial incentive. Both of these processes model the citizenry as relatively passive followers of laws and incentives. Cooperation, on the other hand, is more about intrinsic motivation – it assumes that the cooperators voluntarily work together because they believe that their intrinsic motivations align, and they can achieve more together than separately.

Oddly, research in psychology shows that, to a significant extent, extrinsic motivation actually *inhibits* intrinsic motivation. Lepper [Lepper et al. (1973)] describes an experiment where rewarding children for drawing actually *reduced* their intrinsic motivation to draw.

Part of our prescription in the "Why Can't We" book for making a better world, is to see if we can redesign society to operate more on intrinsic motivation rather than extrinsic motivation. Extrinsic motivation is inherently coercive, because it is always trying to "bribe" people to do things. Intrinsic motivation is the source of creativity and innovation. Intrinsic motivation can be the basis for a more cooperative and productive world.

So the question is this: If we can make our society better by trying to reorient it more towards intrinsic motivation, could AI systems be made better by more attention to intrinsic motivation?

4. Agents, Goals, Subgoals, and Motivation

Now, how do we move these ideas from describing human behavior to describing programs?

We will use vocabulary that is standard in many AI subfields, especially that of planning [Russell and Norvig (2020)].

We model both human users and the top level of programs as *agents*, which we take to be the locus of thinking and action. There may be more than one user, and a program may be organized as more than one agent, as in *multi-agent systems* [Minsky (1986)]. An agent can have *goals*, which are represented by an And/Or tree. That is, each goal can be broken down into *subgoals*. It may be necessary to accomplish all of the subgoals to achieve a goal (And) , or it may only be necessary to achieve one of a set of alternatives to achieve the goal (Or). The tree bottoms out at *actions*. Each action performs a transformation on the state of the world, bringing it to a new state. The agent has an evaluation function that decides whether the action has accomplished the goal or not. The general operation of an agent is to pick one of its goals, and traverse the goal tree attempting to satisfy the goal. This is of course a rough description, and there are many variants, such as probabilistic formulations.

It is the goals (and the corresponding evaluation functions) that provide motivation for an AI agent. In a completely extrinsically motivated system, an external entity like a human programmer or end user, provides the goals and evaluation. Once the goals have been specified, the operation of the program cannot change them. It can break them down into subgoals, it can perform actions that are intended to accomplish the goals, it can report if the goals are or are not satisfied, but that's it.

Intrinsic motivation provides for more dynamic modification of the goal structure. That is, the program itself can add or subtract goals, or re-plan to find a different decomposition of a goal in terms of subgoals.

5. Intrinsic motivation and explainability

Explainability is currently a hot topic in AI, because many popular AI techniques, while they are successful in terms of accomplishing goals, they have trouble providing human-understandable explanations of their results. They sometimes suffer poor acceptance from end users because of this [Gunning and Aha (2019)].

There are several different approaches for what constitute explanations. The program can provide traces of actions; it can display samples of data that are treated in different ways by the program; it can point out which features are most influential; it can break down complex models into simpler ones. These can all be helpful.

But many of these techniques are oriented towards exposing the *what* and the *how*. Programs oriented around extrinsic motivation often have trouble answering questions about the *why*. The only real answer is “because you told me to”. Programs that have some intrinsic motivation can expose their goals, subgoals, and plans as answers to why questions. Users can see if the program's goals align with their own goals.

Some explanation technologies do give some insight into a program's operation, but leave the user at a loss as to what to do if the explanation indicates that the program is not operating as the user would like. They aren't helpful for *debugging* [Lieberman et al. (2018)]. Interacting with the user in terms of motivations and goals provides a more natural means for the user to debug the operation of the program: delete or replace unwanted goals; or modify the motivation mechanism to bring it more into alignment with the user's desires.

6. Examples of AI programs and methods

Let's investigate a few examples of AI techniques and particular AI projects to see how they stack up in terms of their use of intrinsic and extrinsic motivation.

6.1. Reinforcement Learning

Reinforcement learning programs are the quintessential examples of extrinsically motivated AI systems. A reinforcement learning program is driven by a reward function, an external signal which evaluates each state of the environment in which it is learning. At each step, the program can choose from a set of actions, and the chosen action causes a state transition to the next state. The reinforcement learning program learns to choose those actions which maximize the reward. Basically, RL is extrinsically motivated because there is no other motivation than the reward.

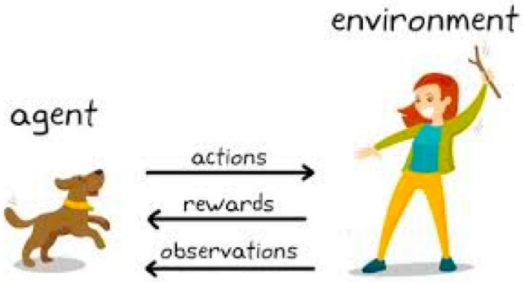


Figure 2: Reinforcement Learning.

Thus maximization of the reward is the highest level goal, and cannot be changed. New goals cannot be generated. Typically, the learner cannot affect the reward function, which remains the same throughout. The learner can, though, internally make tradeoffs like the classic exploration vs. exploitation tradeoff, which you could consider subgoals of the higher goal of learning the reward function. Actions can be chosen by the learner not only to maximize the reward in the short term, but to learn more about the reward function by experimenting. They forgo some amount of short-term reward in the hopes that better knowledge about the reward function will pay off in the long term.

Interestingly, [Barto (2004)] use this move to talk about what they call *intrinsic reinforcement learning*. They apply the reinforcement learning idea recursively, splitting the agent itself into an internal (meta-)agent, and an internal environment. You could think of the internal environment as being the agent's mental model of the world (as opposed to the actual world, which is modeled by the original, external environment). Then, you could think of the agent's using the exploration arm of the exploration vs. exploitation tradeoff, as a kind of curiosity, the motivation for exploration.

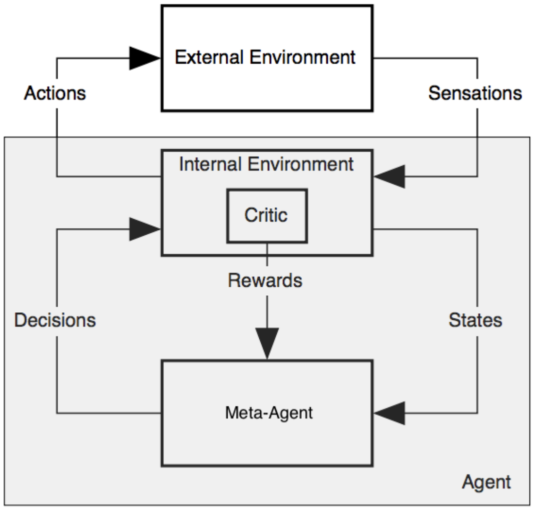


Figure 3: Intrinsic Reinforcement Learning. The RL Agent is split up into an internal Environment and a Meta-Agent

Some RL purists might call that cheating, in that it undermines the simplicity that the original reinforcement learning paradigm was meant to achieve. But it shows that intrinisic motivation has its advantages, even in a learning paradigm that was designed with extrinisic motivation in mind.

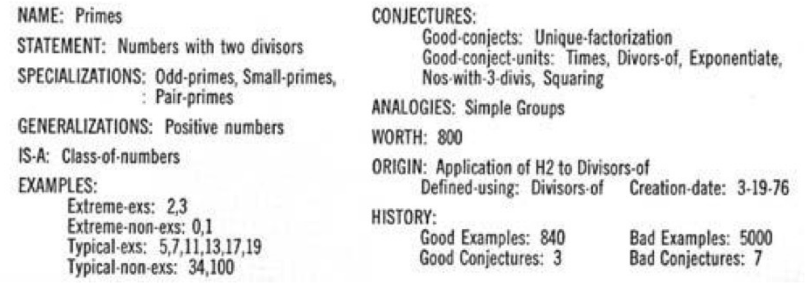
Other learning techniques fall in different places on the spectrum. *active learning*, for example, is more like intrinsic motivation, because the program can decide to see information by choosing goals it thinks will help improve its function.

6.2. AM and Eurisko

At the other end of the spectrum, we have programs like Doug Lenat's AM and Eurisko [Lenat and Brown (1984)]. AM was an "automated mathematician" that attempted to prove new mathematical theorems by search through a space of mathematical concepts, operations, and proof procedure steps. It wound up discovering some classic theorems that it was not given to begin with, and even discovered some minor novelties. Eurisko was a follow-on program that extended AM's capabilities to new domains other than mathematics, and demonstrated success in playing certain video games.

AM wasn't designed to solve a specific problem, but rather to generate "interesting" new theorems. A key capability was that, like a human mathematician, it had the capability of making conjectures. It had the capability of generating new goals on-the-fly. There wasn't any kind of external "reward" signal to direct its activities.

As it attempted to prove theorems, it kept track of which heuristics were more fruitful than others, and in what circumstances. This allowed it to do a best first search which improved its problem-solving capability. It also allowed it to generate new conjectures by



NAME: Primes

STATEMENT: Numbers with two divisors

SPECIALIZATIONS: Odd-primes, Small-primes,
: Pair-primes

GENERALIZATIONS: Positive numbers

IS-A: Class-of-numbers

EXAMPLES:
Extreme-exs: 2,3
Extreme-non-exs: 0,1
Typical-exs: 5,7,11,13,17,19
Typical-non-exs: 34,100

CONJECTURES:
Good-conjects: Unique-factorization
Good-conject-units: Times, Divors-of, Exponentiate,
Nos-with-3-divis, Squaring

ANALOGIES: Simple Groups

WORTH: 800

ORIGIN: Application of H2 to Divisors-of
Defined-using: Divisors-of Creation-date: 3-19-76

HISTORY:
Good Examples: 840 Bad Examples: 5000
Good Conjectures: 3 Bad Conjectures: 7

Figure 4: Lenat's Automated Mathematician.

using analogical reasoning to imagine what its best past heuristics might generate in the current situation, or a generalization of the current situation. It then set out to prove the conjecture.

6.3. Leela

Leela AI [Minsky et al. (2020)] is a learning program for a grid-world environment that combines language, perception and action. It is intended to model how very young children develop concepts of objects, motion and action, and learn language through correlation with perception and action.

It is based on the schema concept of Drescher [Drescher (1991)], in turn inspired by the schemas of Piaget, used to explain early child development. Schemas are like if-then rules, but also include a description of the result of applying the action specified by the then part. Learning proceeds by prediction and experimentation, leading to the creation of new schemas. This follows Piaget's two phase process of assimilation and accommodation. If a prediction is verified, its initial conditions may be generalized. If the prediction is refuted, they may have to be specialized.

Leela is also a good example of an intrinsically motivated AI program. Like AM, it has the ability to make conjectures, through newly generated schemas, which in turn may be confirmed or refuted by experience.

6.4. Lensing

Lensing [Dinakar (2017)] is a mixed-initiative technique for representing perspective in machine learning. Its mixed-initiative character gives it a mix between intrinsic and extrinsic motivation.

It is seldom appreciated in machine learning, but there is a big difference between learning from data that comes from measurements, such as medical instrumentation, and data that comes from people. If you have data from medical instruments, the instruments don't get a vote. The data is objective, although there may be errors or inaccuracy.

The people do get a vote. Each person has their own point of view and perspective. Different people may view the same situation differently. When machine learning aggregates data coming from different people, it is like putting all the data into a blender, and

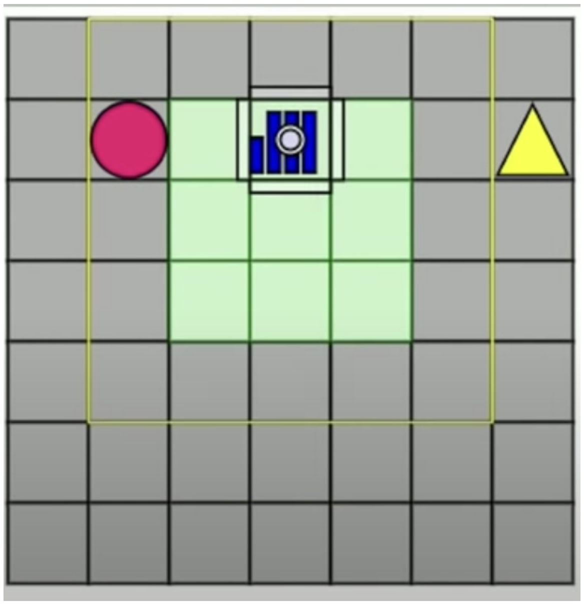


Figure 5: The Leela grid world. The highlighted region is the fovea of attention, containing the user's hand.

pressing the pure button. You might get a tasty smoothie, but the characteristics of the individual components will be lost. The result will be like some sort of average of the various perspectives involved. When the perspectives differ substantially, the aggregate result may satisfy no one, and important insights may be lost.

Lensing makes perspective a first-class object. It is a technique that can be used in conjunction with a variety of different standard machine learning paradigms. A single run of a dataset can generate some predictions and generalizations, which can then be fed back to a person, whom we shall call the *informant*.

Informants are different than annotators, people who may have supplied labels for elements of the original data, if we are using a supervised learning paradigm. The job of the informant is to provide samples of important feedback on the generalizations and categorizations produced by the learning problem, given their point of view. It is not their job to re-annotate the entire dataset, which would usually be impractical.

From the informant's feedback, we compute a lens, a representation of the difference between the informant's point of view, and that of the original annotators, or of the program itself in the case of an unsupervised approach. The lens can be fed back to influence the learning procedure, and another round can commence.

For example, in a medical setting, the original data or labels may come directly from patients, but the informants would be medical professionals. Because their time and expertise is valuable, lensing acts to amplify the impact of their input. The first-class nature of the representation of the lens brings up intriguing possibilities that we could apply multiple lenses to the same data, view different data with the same lenses, etc.

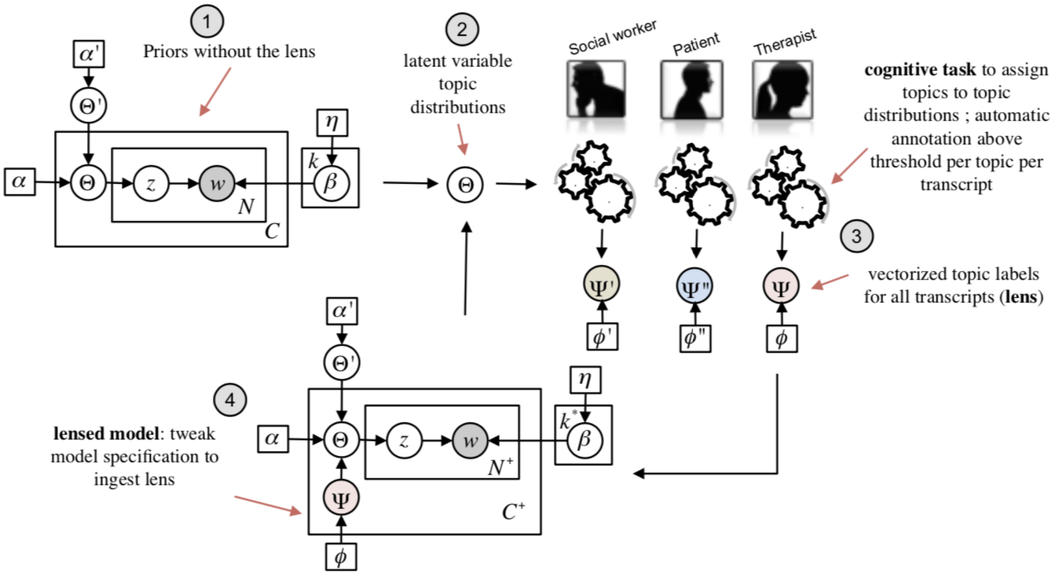

1
Priors without the lens
α'
Θ'
α
Θ
z
w
N
C
η
k
β
2
latent variable topic distributions
Θ
Social worker
Patient
Therapist
cognitive task to assign topics to topic distributions ; automatic annotation above threshold per topic per transcript
3
Ψ'
Ψ"
Ψ
φ'
φ"
φ
vectorized topic labels for all transcripts (**lens**)
4
lensed model: tweak model specification to ingest lens
α'
Θ'
α
Θ
z
w
N+
C+
η
k*
β
Ψ
φ

Figure 6: Lensing. Top left, topic modeling using LDA. Top right, user feedback creates a lens. Bottom, lens is incorporated into the next round of the model.

Lensing provides a methodology for integrating intrinsic and extrinsic motivation. The underlying machine learning technique, and the data itself, give the program an internally motivating factor. But the feedback from the informant makes sure the human and machine motivations are in alignment. Successive iterations of human and machine feedback assure that the system can adapt to changing conditions and changing needs of its users.

7. Conclusion

Understanding the sources of motivation for both human and machine intelligent systems is a powerful tool for designing better systems, and ultimately, for fostering human-machine cooperation. The distinction between intrinsic and extrinsic motivation is a fundamental one. Like yin and yang, they are complimentary, and neither is best in all circumstances. I believe that many current systems, both human organizations and AI architectures, place too much emphasis on extrinsic motivation, and haven't fully explored the power of intrinsic motivation. But you'll make up your own mind about that.

Acknowledgments

Thanks to Henry Minsky, Cyrus Shaoul, and Milan Minsky for organizing the event. Thanks to the MIT CSAIL Genesis Group, and to Christopher Fry, for important ideas and collaboration. This work was funded by the US Air Force Office of Scientific Research and the US Defense Advanced Research Projects Agency.

References

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

A. Barto. Intrinsically motivated learning of hierarchical collections of skills. In *3rd International Conference on Development and Learning*, 2004.

Karthik Dinakar. *Lensing Machines: Representing Perspective in Machine Learning*. MIT, 2017. URL https://dspace.mit.edu/handle/1721.1/112523.

Gary Drescher. *Made Up Minds*. MIT Press, 1991.

Christopher Fry and Henry Lieberman. *Why Can't We All Just Get Along*. http://www.whycantwe.org/, 2018.

David Gunning and David Aha. Darpa's explainable ai program. *Artificial intelligence Magazine*, 40(2):44–58, 2019.

Douglas B Lenat and John Seely Brown. Why am and eurisko appear to work. *Artificial intelligence*, 23(3):269–294, 1984.

Mark Lepper, David Greene, and Richard Nesbitt. Undermining children's intrinsic interest with extrinsic reward: A test of the overjustification hypothesis. *Journal of Personality and Social Psychology*, 28(1):129–137, 1973.

Henry Lieberman, Valeria Staneva, and Yen-Ling Kuo. Debugging probabilistic programming: Lessons from debugging research. In *1st International Conference on Probabilistic Programming*, 2018.

Henry Minsky, C. Shaoul, M. Minsky, S. Kommrusch, and G. Drescher. Self-supervised learning of concepts of physical objects. In *International Workshop on Self-Supervised Learning*, 2020.

Marvin Minsky. *The Society of Mind*. Simon and Schuster, 1986.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2020.

Robert Wright. *Nonzero: The Logic of Human Destiny*. Basic Books, 1994.