Date: Jan 4, 1994
Negative Expertise
Marvin Minsky

Published as "Negative Expertise," International Journal of Expert Systems, 1994, Vol. 7, No. 1, pp. 13-19.

Abstract: We tend to think of knowledge in positive terms -- and of experts as people who know what to do. But a 'negative' way to seem competent is, simply, never to make mistakes. How much of what we learn to do -- and learn to think -- is of this other variety? It is hard to tell, experimentally, because knowledge about what not to do never appears in behavior. And it is also difficult to assess, psychologically, because many of the judgments that we traditionally regard as positive -- such as beauty, humor, pleasure, and decisiveness -- may actually reflect the workings of unconscious double negatives.

==========

"An expert is one who does not have to think. He knows."--Frank Lloyd Wright

Abstract: We tend to think of knowledge in positive terms -- and of experts as people who know what to do. But a 'negative' way to seem competent is, simply, never to make mistakes. How much of what we learn to do -- and learn to think -- is of this other variety? It is hard to tell, experimentally, because knowledge about what not to do never appears in behavior. And it is also difficult to assess, psychologically, because many of the judgments that we traditionally regard as positive -- such as beauty, humor, pleasure, and decisiveness -- may actually reflect the workings of unconscious double negatives.

This inclination is expressed in the "rule-based expert systems" that emerged from research in AI. Virtually all of their knowledge is encoded as positive rules: "IF X happens, DO Y." But this misses much of expertise. Certainly, competence often requires one to know what one must do -- but it also requires you to know what not to do. "IF you are close to a precipice, DON'T walk toward it." An expert must know both how to achieve goals and how to avoid disasters. Sometimes we can take positive measures against accident -- but mostly we do it by avoiding actions that might cause trouble. This essay argues that much of human knowledge is negative.

And the same applies to thinking, as well. In order to think effectively, we must "know" a good deal about what not to think! Otherwise we get bad ideas -- and also, take too long. This raises a number of theoretical issues:

Why is negative knowledge important?

The world is a dangerous place for life. For example, biologists tell us that most mutations are deleterious. This because each animal is already near a sort of local optimum (with regard to its local environment) in the space of mutational variants. And near the top of any hill, most steps go down.

But why is each animal close to a local peak? Simply because evolution itself is a learning machine that is engineered to climb hills. All existing animals had ancestors that avoided enough accidents to have descendants, and those ancestors were just the ones that acquired machinery that enabled them to learn to avoid poisons, diseases, predators, competitors, and other dangerous situations. Of course we also evolved to learn positive goals and ways to achieve them; still, to the extent that our world offers more perils than opportunities, our topmost goal must be -- don't get killed! There are many ways to avoid dangers You can escape your enemies by destroying, controlling , or evading them. Perhaps our societies, cultures, and governments themselves originated in negative goals, namely, for protection against the most common causes of accidents.

The evolution of intelligence brought great new opportunities -- but also gave us great new ways to fail. As soon as we were capable of reasoning, we became susceptible to fallacies. As we extended the range of our plans, we fell prone to more intricate kinds of mistakes. As the arts of speech evolved, this increased the risk of infection by more bad ideas from other minds. The mental, as well as the physical world may also contain more bad than good. Of course, communication can also transmit ideas that give immunities to other, good and bad, ideas.

The brain has many specialized agencies, In the later chapters of [SOM] I argue that these must use a variety of different representations. Some agencies might use script-like structures for representing sequential concepts and story-like exemplars. Others may use tree-like data-structures and/or semantic networks for hierarchical classifications and more complex structures; topographical arrangements for representing spatial and haptic situations, production-like collections of rules for efficient execution of procedures, and "trans-frame" like structures for reasoning about causality. A cursory glance at the index to a neurology book shows that the brain includes hundreds of anatomically distinct "regions" and bundles of fibers that interconnect them. Why should a brain have so many ways to do things. I think, because so single scheme will work for all the many kinds of problems that the world confronts us with. Each problem-solving strategy, each style of thinking, each knowledge-representation scheme -- each works in certain areas, but fails in other domains. Consequently, for each body of knowledge we accumulate, we also need knowledge about when to use that knowledge-base and when to not.

Some of my colleagues have argued that the brain is not a suitable basis for such a

discussion, because it was never really designed to think. Surely (they say) most of that complexity could be avoided when we design such machines from scratch. Surely (some of them maintain) we can construct a single uniform, consistent, and effective logical systems to perform all kinds of commonsense reasoning. I doubt [1991 Minsky] this will be feasible, because consistency and effectiveness may well be incompatible. Other colleagues maintain that we should be able to construct large, uniform neural networks that can learn to do all that minds might need. I do not see much hope of this, because of fear that any very large such network would be prone to accumulate too many interconnections and become paralyzed by oscillations or instabilities. How could we stabilize such systems ? My answer is that one might have to provide a variety of alternative sub-systems, decoupled enough that if each part should fail from time to time, the rest could continue to function so that not all the system will all fail at once. This means that those parts must be suitably insulated from one another. There has been so little recognition of this problem in modern AI that perhaps we need a new term for it. Perhaps we need to breed researchers who can call themselves "Insulationists".

Of course, some insulationist functions already are encompassed by traditional learning theories. Because the most popular forms of neural networks and fuzzy logic can reduce as well as increase their weights, this could tend locally to eliminate 'detrimental' connections. But it is my feeling that although that sort of thing is formally possible, it is heuristically impractical. Instead, I maintain, effective systems will need to be provided with suitable architectures from the start. Perhaps we'll have to design each agency with appropriately engineered machinery to prevent our machines from getting stuck. For example, in [1985 Minsky, Ch.10] we proposed that a typical agency might be built to incorporate Seymour Papert's "Exclusion Principle,", so that when an agency develops a serious internal conflict, the subagents involved should be inhibited so that others can take over.

How much of human knowledge is negative?

We spend our lives at learning things, yet always find exceptions and mistakes. Certainty seems always out of reach. Except in worlds we invent for ourselves (such as formal systems of logic and mathematics) we can never be sure our assumptions are right, and must expect eventually to make mistakes and entertain inconsistencies. To keep from being paralyzed, we have to take some risks. But we can reduce the chances of accidents by accumulating two complementary types of knowledge:

We search for 'islands of consistency' within which commonsense reasoning seems safe .
We also work to find and mark the unsafe boundaries of those islands.

Both as cultures and as individuals, we learn to avoid patterns of thought reputed to yield poor results. In civilized communities, appointed guardians post signs to warn about sharp turns, thin ice, and animals that bite. And so do our philosophers, when they report to us their paradox-discoveries - those tales of Liars who admit to lying, and Barbers who shave all who do not shave themselves. These precious lessons teach us about which thoughts we shouldn't think; they are the intellectual counterparts to Freud's emotion-censors. It is interesting how frequently we find logically paradoxically nonsense to be funny, and when we come to jokes, we'll see why this such a humorous character. For when we look closely, we find that most jokes are concerned with taboos, injuries, and other ways of coming to harm - and logical absurdities can also potentially lead to harm.

I think we have neglected that second aspect -- of asking how experts manage to discern and defend the margins their islands of consistency. It is so hard to study what minds do /i[not] think that this may have placed that subject beyond the bounds of behaviorist psychology, because of its non-behavioral character. And introspective methods also fail because (like most of learning and reasoning) such processes are hidden from consciousness. Neil Agnew pointed out to me that this poses a problem for knowledge engineers -- those who would encode an expert's expertise. Presumably, experts have more effective censors than the rest of us -- but we can't rely upon their introspection to detect the work of their inhibitory agencies. Worse, perhaps as Freud proposed, our censors actively resist their exposure. Still, sometimes outsiders can see what in ourselves what we cannot, by noticing such nuances of behavior as avoiding, forgetting, displaying of temper, rationalizing, or citing only positive instances, etc. [See also Agnew & Brown (1986) for a discussion of confidence in reasoning.]

How can we implement negative knowledge?

One way is to divide the mind into parts that can monitor one another. For example, imagine a brain that consists of two parts, A and B. Connect the A-brain's inputs and outputs to the real world - so it can sense what happens there. But don't connect the B-brain to the outer world at all; instead, connect it so that the A-brain /i(is) the B-brain's world! Then A can see and act upon what happens in the outside world. On the other hand, B can only "see" and influence what happens inside A. This could be enough to help block some kinds of bad patterns of thinking in A.

If A is not making progress toward its goal, force it to review that goal.
If A seems to be repeating itself, make it stop and try something else.
If A does something B considers good, reinforce A's learning system.

If A is occupied with too much detail, then make it take a higher level view.
If A is not being specific enough, then make it focus on more details.
If A appears to be making things worse, suppress it in favor of another agency.
If A asks more than three 'whys' in a row, shift to another agency.

This sort of thing could be a step toward a more "reflective" mind-society. A B-brain could experiment with its A- brain, just as the A-brain can experiment with the real-world objects and the people that surround it. And just as A can try to predict and control what happens outside, B can try to predict and control what A will do. And even though B may have no concept of what A's activities mean in relation to the outer world, it is still possible for B to be useful in the sort of way that a counselor or management consultant can assess a client's mental strategy without having to understand all the precise details of that client's profession.

Emotions and NegExpertise

Negative knowledge is involves in many of the forms of thinking that we term 'emotional', notably those involved with humor, shame, fearful, and aesthetic appreciation. This machinery includes a variety of suppressors, critics, and inhibitors, some of which can inhibit not merely actions but entire strategies of thought. Thus, once one begins to look for it, one finds examples of negative knowledge in many activities that we usually see as positive. In the earliest theories about AI, for example, we emphasized the importance of heuristics for generating efficient search trees. This can be done either by pruning initially larger trees or by suppressing those branches right form the start -- that is, by not thinking of them in the first place. When you decide to leave a room, you don't even think of jumping out the window. Thus, a positive system forces us to generate and test, whereas a negative-based system could more efficiently shape the search space from the start. To do this efficiently, we would have to invent ways to compile each new search generator, perhaps on the basis of previously learned negative prototypes. To wait for inhibition during run time would consume more time.

This relates to what is commonly called creativity. It annoys me how frequently people suggest that the 'secret' of making creative machines might lie in providing some sort of random or chaotic kind of search generator. Nonsense! Certainly, there must be a source of variation -- but that can be supplied by all sorts of algorithmic generators. What distinguishes the performance of a 'smart' or 'creative' artist or problem-solver is not how many trials precede a success, but how few. So the secret lies not in disorderly search, but in pre-shaping the search space so as reduce the numbers of useless attempts. Of course, that's not the whole story. In order to establish an individuality, a creative modern artist must also generate some unconventional alternatives. Doing that may also involve un-suppressing some conventional censors. In any case, from the negative knowledge point of view, we might argue that often beauty is neither in the eye, nor even in the mind of the observer, but precisely the opposite: it may lie in the power to inactivate many of that observer's internal critics.

To explain this, let's consider the role of emotions in thought. It seems generally agreed upon that, on the whole, the positive emotions involve learning what to do, while the negative ones involve learning what not to do. But if so, then I suspect that many emotions that we normally see as 'positive' are actually not. For example, it seems to me that much of our celebrated sense of Beauty may be negative, no matter that we see it as positive. For when possessed by that emotion, many people seem to me to have suspended much of their normal question-asking machinery. When a person says, "how perfectly beautiful this is," they seem also to be saying, 'it is time to stop evaluating, selecting, and criticizing.' They often regard as hostile, requests to be asked to explain why they are attracted to it..

Humor is also usually seen as positive, no matter that the force of a joke is to say, "Don't even think about doing X," or "Don't take it seriously!" Most people are quite unaware that jokes are usually about things that one should not do, because they are prohibited, disgusting, or simply stupid.

Similarly, we tend to think of decision-making as positive. Yet the act of decision, which we often describe as an "act" of free will, is more of a NegAct by nature, because what seems consciously to be the moment of 'making' the decision is actually the moment of terminating the process of considering alternatives.

Perhaps it is the feeling of Pleasure that we consider most positive of all. Yet once we start to see the mind as not one, but a society of processes, then the most extreme pleasure can be seen instead as most negative. For it may mean merely that a certain process has seized control, and has managed to turn off most of the rest. Naturally, that makes it hard to think about anything else. Surely the most extreme form of the control of mental agencies can be seen in what we call mystical experience. For when this happens to a mind, it is like saying to oneself, "Now my problems are all solved. I know the Truth, and know that there is no need to question it, or seek confirming evidence. Stop thinking now, and let all Critics cease. "

We normally think of beauty, humor, pleasure, and decisiveness as positive; is it then paradoxical to claim the opposite? No, not at all -- because we're dealing with things complex enough to constitute 'double negatives'. Putting something in a folder labeled 'negative' can't keep it there, because we then can re-enclose it in a second sign-changing shell! Thus pleasure can seem positive to the agency now in control -- no matter that your other agencies are suffering under its yoke. Thus, enjoying something very much can mean that you've engaged machinery that (i) makes you think even more about that something and (ii) keeps you from thinking of other things.

Conclusion

We tend to think of knowledge in positive terms -- and of experts as people who know what to do. But a 'negative' way to seem competent is, simply, never to make mistakes. How much of what we learn to do -- and learn to think -- is of this other variety? How much of human competence is knowing methods for solving problems, and how much of it is knowing how to intercept and interdict unproductive lines of thought? It is hard to assess the importance of these, experimentally, because knowledge about what not to do never appears in behavior. And it is also difficult to assess them psychologically, because many of the feelings and judgments that we traditionally regard as positive may result from forms of censorship of other ideas, inhibition of competing activities, or suppression of more ambitious goals. It is possible that the importance of this subject itself tends rarely to be recognized, precisely because of the mass of inhibitory machinery that constitutes it. Could it be that our accumulations of counterexamples are larger and more powerful than our collections of instances and examples? Could it be that we learn more from negative rather than from positive reinforcement? Our hedonistic culture holds that learning works best when it seems pleasant and enjoyable -- but that discounts the value of experiencing frustrations, failures and disappointments, either in actuality or in the vicarious forms of forewarnings and admonishments.

References

Minsky, M. (1986) The Society of Mind, New York: Simon and Schuster

Minsky, M. (1991) "Society of Mind: A Response to Four Reviews." Artificial Intelligence, April 1991, Vol. 48, pages 371-396.

Agnew, N. Mck. & Brown, J.L. (1986). Bounded Rationality: Fallible decisions in unbounded decision space. Behavioral Science, 31, 148-161

Several paragraphs of this text are adapted from sections of "The Society of Mind."