the human role in decision making

Josh Schoenwald
PSY 815 / WI 2004
02.26.04

Introduction

The human brain is perhaps the most complex structure known to man, and because it defines our existence it only makes sense that we should strive to understand it. The benefits would seem obvious -- if we could only figure out how people think then we could build artificial systems to mimic our behavior, while maintaining complete control. Understanding human decision making has traditionally followed this paradigm, pursuing an ultimate enlightenment whereby "human" decisions can be made by automated systems, thereby saving manpower, costs, and all the other "problems" associated with humans. However, this 'utopia' is a fantasy that we will never reach, nor should we continue trying. It is my belief that automated systems are tremendously important to our society, but only to the degree that they should be our assistants, not our replacements. Taking the human out of the equation is a nearly impossible task, and one that would leave us at the hapless mercy of our creations should we succeed. In this paper I will critique the traditional Bayesian, heuristic, and regression models of behavior, and suggest new paradigms that get to the heart of what it means to understand human decision making.

Some traditional behavioral models evolved out of the idea that humans, being intelligent creatures who pursue what's in their best interest, will always strive for what we believe will give us the greatest benefit. Expected utility theory operates with the notion that we survey our options and choose the one that maximizes utility. Mathematical models (Bernoulli, 1954) have been drawn up to assess this and predict how people should respond. In other cases, Bayesian networks (Morawski, 1989) attempt to use the notion of probability theory to underlie what people believe. Similarly, heuristic approaches attempt to model ways in which humans typically behave under certain conditions, such as the gambler's fallacy (Schulman, 2002) and base rate neglect (Kahneman, Slovic, & Tversky, 1982). Mathematical regression models take the idea of weights into account (Dawes, 1979) but still fail on many accounts to thoroughly understand decision making.

Bayesian Networks

Bayesian networks rely on probabilities of factors to determine the "best" decision. A simple example might go like this: you know that the chances of having a certain disease are about 50% in a given population. There is a test which is 80% accurate when it gives a positive or negative response. Given that the test comes out positive, you can infer that you have an 80% chance of actually having the disease. But what if the chances of someone in the population having the disease are only 2%? Given a positive result from the same test which is 80% accurate, the actual probability that you have the disease is now still only 7.5%. While these methods might be a great way to quantify some partially ambiguous situations, they are very limited in their scope and cannot be expected to account for human behavior in real world situations. While these models do not represent the latest in related research, they share the basic foundations and goals, namely that continuing theoretical iterations will produce some definitive theory.

First, knowing the prior probabilities is often not a realistic scenario. While betting on sports games is rarely done without prior odds, practical situations often do not have this luxury. Furthermore, even knowing the probabilities leads you to question the source; Bayesian networks make the assumption that not only do you have access to the prior probabilities, but that those probabilities are completely accurate and unbiased. But even if we can assume that we know the prior probabilities and that these probabilities are unbiased, Bayesian networks also assume that there are no other constraints acting on the system, and if there are, that these constraints be clearly defined. For instance, back to our disease example; in addition to the 2% prior probability of having the disease, and the 80% accuracy of the test, what if we assume that the disease is susceptible to a certain protein found in beef? In order to use a Bayesian Belief Network to calculate the probability of having the disease, you would need to know how much beef one had consumed in whatever time frame relevant to the vulnerability of the disease. Of course this is only one example but it illustrates the point that you cannot account for all the variability that is likely to have an effect on believability as it applies to the human decision maker. Not only are there too many factors (or potential factors) to account for, but the world and the people in it are in a state of constant change, and no (relatively) static model can be expected to constantly maintain an accurate representation.

As a result, Bayesian networks can be useful in some situations to understand some simple dynamics or to get a feel for trends. Often when first learning a new environment, it's helpful to know some general trend information. For instance a new car

salesman might be taught to go easy on potential car buyers late in the day, because most of them are tired from the day's work and less likely to engage in negotiations. Obviously, it is easy to understand why you wouldn't want to make a steady rule about negotiating tactics based solely on time of day. The reality is that real world situations don't follow all the rules of formal modeling, (no matter how many variables) and should not be sought after as replacements for human agents.

Heuristic Models

Heuristic approaches to decision modeling are slightly more appealing in terms of their methods. Scientists such as Kahneman and Tversky have proposed sets of heuristics which attempt to explain human decision patterns on various situations. While the crux of this methodology is correct in how it identifies the trends, it fails to capture true motivations for observed actions. Referring to them as instances of the sample size or base rate neglect phenomenon are trivializations and often arise from lab experiments with little applicability in real world situations. A typical example might look like this: Suppose there are two coins on a table, and each coin is flipped 6 times. The results are as follows:

coin 1: H, T, H, T, T, H coin 2: T, T, T, T, T, T

Statistically, the probabilities of each set occurring are the same, namely 1.6%. This trend would probably be considered an instance of a 'misperception of a random sequence.' While it might be true that most people would fail this test, the test is misleading. People's natural ideas of randomness tend to focus on random products, rather than random sets of products (Nickerson, 2002). In other words, people can recognize that the chances of getting the exact sequence of head and tails as represented by coin 1 and coin 2 above are the same, however randomness is typically defined not by the product itself, but on the set of products. In that case, using a random sampling we would expect (even with a small sample size of 6) that there would be about 3 heads and 3 tails. Even if we allow some flexibility, knowing that small sample sizes often do not represent the mean of larger sample sizes, getting 6 tails in a row would seem unlikely, and not random.

Examples like this are trivializations in that they do not accurately represent the complexity of the world. In the coin example, perhaps there is no further complexity to consider, however the utility of such methods to figure out the probability of getting one coin sequence over another are lost in the simplicity of the task -- it would take too long to calculate to be useful. The mistake made by many theorists is in assuming that trends observed in fictional coin experiments are relatable to real world complex-

ities. I believe this is a major problem with many laboratory experiments (as opposed to naturalistic (Hutchins, 1995) or staged world studies (Woods, 1993).

These rule-based explanations of behavior are typically very good at describing the trends that emerge as people are asked to judge the likeliness of flipping heads and tails, of boys and girls being born in a certain order, or of a certain personality description as being reflective of a lawyer or engineer (Edwards & von Winkerfeldt, 2000). The problem with these studies is that they lack a true value in situations where such decisions are important. To take these experiments and findings directly into the real world is akin to asking FDR to decide whether to prepare Pearl Harbor for an attack based solely on the occurrences of previous attacks. This example may seem extreme, but it underlies the notion that when we have choices to make, there are *always* other factors to help us make this decision. This is true especially when the stakes are highest. High stakes situations force us to be more cautious in our decisions since the consequences for being wrong are too great to bear. But even in mild situations such as deciding which route to take to the movie theatre are likely to employ as many information channels as readily available. For instance, you might take into consideration construction and traffic patterns, but you probably wouldn't take the time to determine the real-time average speed of traffic on each route and then choose the quickest.

Another factor that is often extremely significant in decision making that cannot be accurately accounted for in heuristic modeling is that of time. Clearly, the impact of the notion of time is a very subjective one across people, and the concept of having "enough time" vs. "too little time" will vary wildly from person to person and even with the same person from situation to situation. Thus, modeling time constraints as a function of heuristic practices is difficult. However, time is constantly playing a role in the decisions we make. While it's true that many decisions are relatively time-independent, such as which type of detergent to use, most decisions must be made in the context of some situation that will not last indefinitely. Many times, the severity of the particular situation is directly tied to time. Intuitively, we can appreciate that people will make different choices given the same scenario if the time constraints are severely distorted. Some would argue that this defines an area where automated mechanisms, employing some heuristic method of decision analysis, could make a 'better' decision in a short amount of time. Because automated systems can evaluate parameters faster than humans in most situations, they could make an informed decision while considering tens or even hundreds of factors when a human might take an exponentially longer time to make the same decision of told to consider the same factors.

The problem with this evaluation is that it trivializes the difficulty of defining these factors explicitly and exhaustively. Additionally, the world is not static, so even if we could define all the appropriate factors to make a decision in a certain context, the scenario could change completely. Take for instance a military advance on an entrenchment of enemy troops. Suppose all available evidence tells us the enemy is a company of 150, holding their position. Our forces of 500 would likely continue to advance and attempt to take their position. But what if we spy a 1200-man battalion lying just beyond the company? Now our goals have changed from taking their position to retreating unnoticed to wait for additional troops. Being bound or merely prompted by the tenets of some automated system would likely lead to a false sense of security because we tend to believe that the automated system's suggestions or actions are correct; even in the face of reliable contradictory information (Skitka et al., 1999; Rasmussen 1986; Vicente & Rasmussen 1992). Because new evidence may come in any form and with myriad repercussions, it is impossible to model all scenarios that might have a significant impact on decision making.

Regression Models

Models of regression have also been used in attempts to model human decision making behavior. These models employ a number of weights attributed to certain identified factors that place an importance on those weights. For instance, in deciding which graduate students should be admitted to a program, it might be determined (through extensive analysis of prior records) that GPA is a far weaker indicator of eventual success than is the combined effects of years removed from college and years of industry work. While the ability to weight different factors is important, it is not sufficient to account for the complexities of real world decisions. Additionally, it does not solve the problem of quantifying variables which are inherently unquantifiable. Robin Dawes makes the distinction that in the process of admitting graduate students, the goal is to find candidates by predicting some variable termed "self actualization." Dawes acknowledges that this term is not easily quantifiable, but makes the assumption that this variable (which has yet to be defined) is positively related to "intelligence, to past accomplishments, and to ability to snow one's colleagues. In our applicant's files, GRE scores assess the first variable, undergraduate GPA, the second, and letters of recommendation, the third" (Dawes, 1979). I think one would have a difficult time arguing that GRE scores are unilaterally indicative of intelligence, or that undergraduate GPA is the only past accomplishment relevant to graduate applicants, or even that letters of recommendation represent a consensus opinion. However, even if those held true; you cannot explain justifications for a novel term with no definition. The fact that the regression model derived from the convenient

use of these three indicators can choose for admission applicants who maximize these criteria should come as no surprise. That this method is automatically thought to indicate a high level of self actualization is absurd. Omitting variables such as prior work experience, research interests, past colleagues, heritage, faculty research projects, etc. would be a foolish way to pick a graduate student, and I'm sure all graduate committees use more than 3 steadfast, simple criteria to determine admissions for the majority of applicants; allowing of course for those students who fall far above and far below some generally-accepted ranges.

The problem of knowing what to look at is a very important issue here, and one that regression models cannot effectively take into account (Dawes, 1979). The notion however that experts are good at quantifying these variables, just not integrating them is still subject to the variability of applicants in all categories from year to year, not to mention new categories that may spring up or old categories that might be outmoded. One ought to ask if the process of creating these linear models year in and year out is worth the convenience of not having to hand-pick graduate students. Is this really lessening the work or just changing it?

Alternate Paradigms

The concept of making decisions is a uniquely human one. Decisions represent junctions at which one course of action is taken over another (or more than one), even if the other courses of action are not explicitly defined. The ability to recognize novel action paths and to re-interpret data in light of new information is crucial for operating in real-world environments. Therefore, eliminating the human agent from the decision equation is almost impossible. Three main themes which I believe underscore this idea are: 1: humans have goals, not machines; 2: humans are context-sensitive while machines are literal-minded; and 3: new technology only changes requisite expertise, it does not eliminate it.

The first theme highlights the idea that humans are goal-driven. Every thing we do is driven by multiple goals (Chow et al., 2000). These goals continuously interact at varying levels to direct the thrust of our behavior. For example, assume your 'goal' is to go buy milk at the store. While this may be your most important goal, to come home with milk, there are subgoals such as not getting in an accident while driving, not getting caught in excessive traffic, not paying more than about $2 for the milk, etc. While these goals may usually lie dormant in that you don't think about them explicitly, they are nonetheless important. If you should get a phone call that your mother has gone into the hospital, all of a sudden getting the milk probably isn't your biggest priority.

Automated systems do not have goals. Machines do not strive to achieve some level of satisfaction or accomplishment. They are designed by humans to help us in reaching *our* goals. There's little doubt Neil Armstrong wouldn't have uttered those famous words without the aid of technology and computer systems. But we would never consider the computers as desiring to reach the cosmos. This notion however is surprisingly prevalent in our society. Consumers, designers, and engineers alike talk about what machines "do" as if they have a mind of their own. Automated systems are so complex that they have a perceived animacy which makes us think they have human qualities (Sarter & Woods, 1994). One of those personifications is believing that machines carry their own goals -- independent from human goals. Because these systems to not have their own goals and self-serving ambitions, they will never be suitable replacements for humans, who will always have them.

The second theme identifies the nature of the computer-human relationship. Although some Artificial Intelligence scientists might disagree, there is a fundamental contrast between humans and machines. Namely, humans are context-sensitive beings while machines are literal minded (Weiner, 1950). This dichotomy, dubbed Norbert's Contrast, outlines the inherent limits of automated systems. They are not optimized for, nor are they good at taking relevant context into consideration when carrying out their functions. This is not a limitation of computing power; the problem is knowing what to look for. Using experts to pinpoint and map the relevant factors (Dawes, 1979) is not an adequate method of overcoming this limitation. While experts are good at understanding the domain and can often identify factors that would be relevant in most situations, they are notoriously bad at externalizing such information (Klein, 1998). Furthermore the only scenarios where these expert-defined factors seem exhaustive is against a list of other factors devised by some experimenter with a limited knowledge of the domain and who is conducting an experiment in the laboratory -- not in a natural setting. Even though domain specifics may be used, conducting an experiment in a lab is not the same as observing the same behavior as it naturally occurs (Klein, 1998).

The third theme points out another related myth about computers and humans; namely that computers are stakeholders. Similarly to how automated systems do not have their own goals; they also are not stakeholders. If a plane crashes into a mountain, there are no families of robots mourning the loss of their beloved autopilot. Machines exist to help us humans, who are always the stakeholders in any system. This fact illustrates the inherent danger in creating automated systems which attempt to assume ever-increasing levels of control without assuming requisite levels of respon-

sibility. Humans will ultimately be the ones responsible, and in order to carry this out they need ultimate authority to act on their own behalf. The authority-responsibility double bind (Cook, 1994) plays out in many contexts, where humans in some situation are responsible for the outcome of a situation without having the necessary authority. One such example is in aviation, where pilots must adhere to the Traffic and Collision Avoidance System (TCAS) warnings and directions from Air Traffic Control. When those directions interfere, the pilot can and usually will be faulted if something goes wrong. If the pilot adheres to the TCAS warning, he's faulted for disobeying ATC, and vice versa of he adheres to the ATC directions. Alleviating this double bind requires the human to have complete command of the situation (Billings, 1991). Adding automated systems that replace human operators do *not* replace the humans as stakeholders.

Some automated system designers and research labs strive to develop automates systems that take the place of humans. Usually the argument is that humans are unpredictable, fallible, and prone to error. Given this premise, it seems obvious that automated systems, which by contrast seem predictable, infallible and relatively immune to error would do well to replace the human operators wherever possible. Additionally, the enormous power of today's computers perpetuates this Sisyphian struggle. The idea that new technology can be a simple substitution for people is an oversimplification called the substitution myth (Woods and Dekker 2002). In reality, adding or expanding the role of automation changes the human's role, it does not eliminate it (Sarter, Woods, & Billings, 1997). Because the human must still be involved at some level–because they are the stakeholders and ultimately responsible for the consequences–adding automation only changes that role, often from operator to supervisor. For instance, introducing a robot that can weld does not get rid of the human. While the human welder might be gone, there is still someone needed to make sure the robotic welder is doing things correctly, and to shut it down if it malfunctions. In order to do this effectively, this supervisor must understand how the robot works and the rules by which it operates. In the case of welding this might be very clear–either the robot welds correctly or it doesn't, and the results are immediately observable–but in other situations, such as computer-controlled pressure regulation, effects of a poorly operating system may not be immediately apparent, or they may manifest in other systems which are working correctly. Providing effective support necessitates more training and, surprise–expertise! It then becomes clear why systems where there is "no knowledge needed" are usually a very bad sign. Not being able to observe the workings of a system does not remove the need for expertise, it only hinders the ability of humans to effectively respond when things don't go as expected.

Conclusion

The increase of technological devices and automated systems hold a fantastic promise for our future. However, this promise will not be realized through the systematic replacement of humans with machines. The power lies in augmenting human performance through an integrated team of computers and humans. Humans are always the stakeholders and the ones who have goals. We are context-sensitive beings who have the power to innovate in novel situations. Machines represent exciting new ways we can work to achieve those goals, yet they are literal minded and cannot be expected to achieve those goals without us. Additionally, in order for the role of technology to support, rather than replace human endeavors, we must recognize that ultimate control must always rest with humans, who always have ultimate responsibility.

References

Bernoulli, D. (1754). Exposition of a new theory of the Measurement of Risk (L. Sommer Trans.). In *Econometrica, 22*, p.23-26 (Original work published 1738).

Billings, C.E. (1991). Human-Centered Aircraft Automation: A Concept and Guidelines. In *NASA Technical Memorandum 103885*. Moffett Field, CA: NASA-Ames Research Center.

Chow, R., Christoffersen, K., & Woods, D.D. (2000). A Model of Communication in Support of Distributed Anomaly Response and Replanning. In *Proceedings of the IEA 2000/HFES 2000 Congress*. Human Factors and Ergonomics Society, July, 2000.

Cook, R.I. & Woods, D.D., (1994). Operating at the sharp end, in Human Error. In Bogner, M.S., Ed. *Medicine*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Dawes, R. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist Vol. 34, No. 7*, p.571-582.

Edwards, W., & von Winterfeldt, D., (2000). On Cognitive Illusions and their Implications. In T. Connolly, H. R. Arkes, and K. R. Hammond (Eds), *Judgement and decision making*. Cambridge University Press. Cambridge, UK.

Hutchins, E. 1995. *Cognition in the Wild.* MIT Press, Cambridge, MA.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgement under uncertainty: Heuristics and Biases* Cambridge University Press, New York

Klein, G. (1998). *Sources of Power*. MIT Press, Cambridge, MA.

Morawski, P. (1989). Understanding Bayesian Belief Networks. *AI Expert, Vol. 4, No. 5*, p.44-48.

Nickerson, R. (2002). The Production and Perception of Randomness. *Psychological Review Vol. 109, No. 2*, p.330–357.

Rasmussen, J. (1986). *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*. North Holland, New York.

Sarter, N.B. & Woods, D.D. (1994) Decomposing Automation: Autonomy, Authority, Observability and Perceived Animacy. *First Automation Technology and Human Performance Conference, April 1994.*

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.) *Handbook of human factors/ergonomics, second edition*. Wiley Press, New York, NY.

Schulman, D. (2002) The gambler's fallacy: why we expect to strike it rich after a losing streak. *Psychology Today, July-August, 2002*

Skitka, L., Mosier, K. L., & Burdick, M. (1999) Does Automation Bias Decision-making? *International Journal of Human-Computer Studies 51*, 991-1006.

Vicente, K.J., & J. Rasmussen (1992). Ecological Interface Design: Theoretical Foundations. *IEEE Transactions on Systems, Man, and Cybernetics, Vol. 22, No. 4*, p.589–606.

Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Doubleday, NY.

Woods, D. D. (1993). Process-tracing methods for the study of cognition outside of the experimental psychology laboratory. In Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C. E. (Eds) *Decision making in action: Models and methods.* Ablex, Norwood, NJ, p.228-251.

Woods, D.D. & Dekker, S. (2000). Anticipating the effects of technological change: a new era of dynamics for human factors. *Theoretical Issues in Ergonomics Science Vol.1, No 3,* p.272-282.

Woods, D. D. & Tinapple, D. (1999). W^3: Watching Human Factors Watch People at Work. *Presidential Address, 43rd Annual Meeting of the Human Factors and Ergonomics Society, September 28, 1999*. Multimedia Production at http://csel.eng.ohio-state.edu/hf99