Thinking About Thinking in Computers, Robots, and People*

Daniel T. Levin, Megan M. Saylor,
D. Alexander Varakin

Psychology and Human Development
Vanderbilt University
Nashville, TN 37209, USA
daniel.t.levin@vanderbilt.edu
m.saylor@vanderbilt.edu
alex.varakin@Vanderbilt.Edu

Stephen M. Gordon, Kazuhiko Kawamura,
D. Mitchell Wilkes

School of Engineering
Vanderbilt University
Nashville, TN 37209, USA
stephen.m.gordon@vanderbilt.edu
kaz.kawamura@vanderbilt.edu
wilkes@vuse.vanderbilt.edu

Abstract - **Although much research has focused on emerging understandings of representation and mind, very little research has explored the adult understanding of different representational systems. The current experiments demonstrate that adults differentiate intentional and nonintentional representational systems when reasoning about them. In the first experiment, subjects were taught a novel feature of either a computer or a person, and asked whether the feature characterized other representational systems. Results indicate that when making inductions from a person to a computer, they do so more for nonintentional mental properties than intentional mental properties. In a second experiment we asked adults to make predictions about the behavior of a person, a computer, and a robot, and found that they assume that humans engage in more object/category related behaviors, and that computer systems engage in more spatial/featural behaviors. Combined, these data establish the existence of a distinct category of intentional mental entities, and give guidelines for its deployment in predicting a variety of representationally mediated behavior.**

Index Terms – Intentions, Concepts, Theory of Mind

I. INTRODUCTION

When people use computers, and interact with robots, they need some understanding of the internal processes inherent to each. Consider, for example, what happens when one of these systems fails or crashes. The user must determine the degree to which the failure was caused by their own behavior, by an internal software problem, by a hardware failure, or even by faulty input information. All of these decisions are, to a degree, predicated on an understanding of the internal workings of these systems. So, how do users think about these processes? They might presume that the representational processes inherent to computers and robots are fundamentally the same as those inherent to people. In such a case, individuals might suitably generalize the folk psychology they have learned for people to these systems. Indeed, drawing this kind of deep similarity between brains and computers as symbol processing representational systems would be consistent with many of the theories that underlie cognitive science. This commonality is also consistent with research demonstrating that people tend to treat computers as social agents [1].

On the other hand, research in cognitive development has emphasized early-emerging distinctions between mechanical and psychological causality that would tend to different patterns of reasoning about computers and people. This research suggests that during the first year of life, infants begin to distinguish goal-directed intentional action from mechanical action [2][3] and have different expectations about the movement of people and the movement of inanimate objects [4][5]. Many researchers argue that this early contrast becomes elaborated into a domain theory that allows older children to understand the representations underlying others' behavior. As they develop this representational "theory of mind" (TOM), children learn that representations are not simply copies of the world, but rather are meaningful interpretations of the world that are intimately related to a person's beliefs, desires, goals, and experiences.

One interesting gap in the TOM research is that adults' understanding of the mind is less well understood. While the assumption seems to be that children's concepts lead to a well-functioning TOM in adulthood, recent research has questioned the degree to which adults have an effective understanding of the representations involved in everyday visual tasks [6]. Thus it is possible that the kind of computer-directed social behaviors observed in previous research reflects either adults' realization that early distinctions might need to be overridden (because they are sometimes inappropriate), or a vague understanding of differences in the representational abilities of people and computers. Alternatively, it is possible that adults retain distinctions between psychological and mechanical representational systems, but do not apply them when executing automatic social behaviors. The current experiments were designed to test concrete predictions about the concepts adults apply to representational systems, and to explore the developmental outcome (at least in the young adults we typically find in college) of early distinctions between representational kinds.

* This material is based upon work supported by the National Science Foundation under Grant No. 0433653.

The experiments in this paper investigate adults' understanding of representational systems using two strategies common in the concept literature. First, in Experiment 1, we used an induction paradigm to test the degree to which subjects would generalize newly-learned properties from one kind of representational system to another. Experiment 2 tested whether the distinctions observed in the first experiment would affect the concrete predictions people make about the behavior of mechanical and psychological agents. In addition to understanding how people reason about the relatively pure cases of computers and people, we were interested in understanding how adults would respond to systems that share properties of both kinds, so we also asked subjects about anthropomorphized robots that combine properties of both kinds.

II. Experiment 1

In our first experiment, we used an induction paradigm to test the degree to which adults would functionally differentiate intentional and nonintentional entites. Subjects responded to a series of scenarios in which novel properties were ascribed to either a personal computer, or a human. These included physical properties, biological/homeostatic properties, intentional mental properties, and nonintentional mental properties. For example, for one of the intentional mental properties, subjects were asked to "Imagine that John's brain attempts to predict what a person is going to do next using an 'M-rule' that takes their specific preferences into account using an algorithm built around a set of Gaussian kernels." The key to this kind of property is that it reflects the need to understand the representational states of people (see Appendix for the complete set of mental properties). That is, these properties reflect understandings of beliefs, desires, and goals. We chose to create scenarios describing novel mental properties such as the one above to test the hypothesis that subjects have expectations about the general kind of process they represent, not the degree to which any specific process or algorithm characterizes human or machine thought.

The nonintentional mental properties generally referred to information processing capacity limits and basic memorial processes that did not involve interacting with others or inferring beliefs. For example, one nonintentional mental scenario asked subjects to "Imagine that John's brain can keep a large amount of information for a long time, and sometimes organizes it by grouping things temporally (e.g. by putting things together that occurred at about the same time) using a regression-based principle." Then, subjects were asked whether another adult, a mouse, a thermostat, a robot, and a computer would also use the same process or have the same property. If subjects differentiate representational systems, they should be less likely to attribute these properties across category boundaries. More interesting, if these functional categories are specifically organized around intentional mental properties, then this effect should be specific to these properties. In this report we focus our analysis on the mental properties and the humans, robots, and computers.

We also manipulated how the robot was described across subjects. For half of the subjects, the robot was given an anthropomorphic name ("OSCAR"), and for the other half it was given a nonanthropomorphic name ("SOCAR"). This contrast produced few consistent results, so it is not considered further in this report.

III. Experiment 1 Method

A. Subjects

A total of 23 (19 female) General Psychology students at Vanderbilt University and Nashville Community College participated in this experiment. Their mean age was 25.9 (SD=8.4).

B. Stimuli and Procedure

A total of 20 induction scenarios were created, with four examples in each of five domains. Each scenario asked subjects to imagine that a novel process or property was characteristic of either a person (named John) or a computer (a Dell personal computer), then to indicate whether the property was characteristic of six other things. When the property was taught about the person, the targets were another person ("Popol", a villager from the Amazon), a mouse, a thermostat, a robot, a Dell personal computer, and an Apple Macintosh personal computer. When the property was taught about the Dell, John replaced the Dell as an induction target. Both robots were briefly described as general-purpose devices. The nonanthropomorphic robot, "SOCAR", was described as an industrial robot, designed to "handle a wide variety of industrial materials", and to "meet as many task objectives as possible with a minimum of waste". The anthropomorphic robot, "OSCAR", was also described as a general purpose robot, but also as having the "goal" of interacting productively with people: "In interacting with people his primary goal is to be as productive as possible and to meet as many possible needs of the person he is interacting with."

The five domains included scenarios testing *homeostatic* process properties (for example, a specific pattern of circulation of fluids throughout the system), *physical* properties (for example, the fact that the material the object is made of makes it turn blue when submerged in water), *"external impact"* properties (for example, an object's response to gravitational forces), *intentional* mental properties (for example, for one of the intentional mental properties, subjects were asked to "Imagine that John's brain attempts to predict what a person is going to do next using an 'M-rule' that takes their specific preferences into account using an algorithm built around a set of Gaussian kernels."), and *nonintentional* mental properties (for example, grouping information in long term memory using a "regression-based principle"). Each subject read two scenarios of each kind that asked about inductions from a person, and two that asked about inductions from a computer. The induction source was counterbalanced across domains.

Subjects responded to the scenarios by completing paper packets individually or in small groups.

IV. Experiment 1 Results

Results supported the basic hypothesis. First, subjects were significantly more likely to assign mental properties taught about one person to another (78%) than they were to assign properties taught about a person to a computer (52%), $t(22)=4.601$, $p=.019$. The same was true of properties taught about the computer (computer -> computer induction: 80%, computer -> person induction, 43%, $t(22)=4.601$, $p<.001$) More important, subjects were significantly less willing to extend intentional mental properties from a person to a computer than they were for nonintentional mental properties. 64% of nonintentional mental properties were assigned from people to computers, whereas only 39% of intentional mental properties were, $t(22)=3.749$, $p<.001$. This effect was not present when making inductions from computers to people: 42% of intentional properties taught of computers were assigned to people in comparison to 45% of nonintentional properties, *ns*.

Inductions to the robot suggest that subjects did not anthropomorphize the robot. Overall, subjects were significantly more likely to assign mental properties taught about the computer to the robot (52%) than when the properties were taught about the person (37%), $t(22)=2.299$, $p=.031$. Subjects assigned 39% of intentional mental properties to the robot when they were taught about a person, and 45% of nonintentional properties when they were taught about a person, $t<1$, ns. When they were taught about a computer, subjects assigned 45% of intentional properties to the robot, and 58% of nonintentional properties to the robot, $t(22)=1.447$, $p=.162$.

Although our focus is not on inductions to the thermostat and mouse, we briefly note that they were not significantly different for intentional and nonintentional mental properties (all *p's*>.10). When the computer was the source, subjects made 20% intentional inductions and 37% nonintentional inductions to the thermometer, and 45% intentional and 32% nonintentional inductions to the mouse. When the person was the source, subjects made 20% intentional inductions and 9% nonintentional inductions to the thermometer, and 52% intentional and 52% nonintentional inductions to the mouse.

V. Experiment 1 Discussion

These results of Experiment 1 suggest that subjects differentiate human and computerized representational systems using an intentional/belief-desire framework. Not only were subjects significantly less likely to assume mental property commonalities between computers and people than they were for members of the same category, but they were significantly less willing to make inductions from people to computers for intentional mental properties than nonintentional mental properties. One interesting finding was that this inductions pattern was not symmetric. Intentional properties had no special status when making inductions from computers to people. One possible reason for this might be that people interpreted the intentional properties as nonintentional when they were characteristic of computers. Future experiments will be designed to follow up on this possibility.

In addition, subjects appeared to anthropomorphize the robot minimally. The relative prevalence of computer -> robot inductions over human -> robot inductions suggests that subjects considered the robot to be more functionally similar to the computer than the person.

VI. Experiment 2

In the second study, we tested the degree to which these different representational theories would affect specific predictions about the behavior of intentional and nonintentional systems. Subjects in this experiment first read descriptions of either a person, a computer, or an anthropomorphized robot, and then made predictions about its/his behavior in a series of six scenarios. We asked subjects to reason about scenarios in which a system's response was either object-directed (as would be the case for an intentional system), or location directed. For example, one of the scenarios was directly based on previous research demonstrating that infants differentiate human action from mechanical action by assuming that mechanical reaches are location-directed whereas intentional reaches are object directed [7]. In this previous experiment, infants were shown a hand or rod reaching for one of two objects (e.g., a bear on the right and the a ball on the left). The objects' locations were switched and infants were shown trials where the hand or rod/claw reached for the old object at a new location, or a new object at the old location. Infants revealed a preference for new object-old location trials for the hand, indicating they were tracking the relationship between the reacher and her goal. In contrast, for the rod/claw they no such preference, indicating they did not have the same expectation for object-directed behavior on the part of the rod. In the present experiment, we adapted this situation, and others like it, to allow subjects to make predictions about the behavior of one of three different entities, a human, a computer, and an anthropomorphized robot. The basic prediction is that predictions about the human will reflect the presumption that it has goals, and uses semantic categories to organize the word, whereas predictions for the computer will reflect a more mechanical form of reasoning in which actions generally do not have goals, and instead are directed at locations.

VII. Experiment 2 Method

A. Subjects

A total of 33 subjects completed the experiment. Of these, 11 completed the Human Condition, 12 completed the Robot condition, and 10 completed the Computer condition.

B. Stimuli and Procedure

Subjects read a series of six scenarios asking them to make predictions about an entity's behavior, or a judgment about the optimal instructions to give it. The first scenario (the "Switch" scenario) described two trials of a "reaching exercise" in which an entity reached for one of two objects in

two locations. Then, the objects were switched to new locations, and the subject was asked whether the entity would reach to the old location (and therefore the new object) or to the new location (and the old object). The intentional response would be the reach for the old object at the new location demonstrating that subjects believe that the entity was engaged in a goal directed reach to the object and would reach for it again. The second scenario (the "Imitate" scenario) described an exercise in which the entity was "imitating" or "repeating" a person's actions (throwing balls into a basket). On the first two trials, the action is successful, and the entity successfully repeats the action. Then, on the third trial the person misses the basket, and the question is whether the entity will imitate the specific action (the miss), or will recognize the goal of the action and throw the ball into the action. In the "Category" scenario, an array of six objects is pictured first in a disorganized state, then organized by category in two different ways. In one, the objects are grouped by perceptual similarity (the dark square objects are grouped), and in another they are organized by semantic category (candy and office supplies). Subjects were asked which of the two organizational schemes the entity would use with the categorical organization putatively characteristic of an intentional system. In the "row" scenario, subjects are told that the entity has "reached" for the first, third, and fifth item in a row including writing utensils, and other similarly shaped objects. The three reached-for items all happen to be writing utensils, and 6th item is a marker, and the 7th a screwdriver. The question is whether the system will continue spatial pattern of reaching and go for the 7th item (the intentional response), or continue reaching for writing utensils and reach for the marker at position 6. In the "description" scenario, subjects were shown a picture of a floppy disk, and a red pen and asked whether it would be better to direct the entity to "lift the red pen" (intentional) or to "lift the object on the left" (Mechanical response). Finally, in the "card" scenario, subjects are shown two pictures, each representing a trial in a card sorting task in which cards with a circle or square on the left or right side are placed into boxes labeled with a matching shape in the matching location. Then, on the critical trial, subjects were asked what the entity will do with a card that matches the shape, but not the location, of the illustration on one box, and the location, but not the shape of the illustration on the other box. The intentional response would be to assume that the system would put the card in the box that matched in shape, not location.

Each subject completed one of three different conditions. In each they made predictions about a different entity. In the Human condition, subjects made predictions about a man named "John" (a picture of a college age male was provided as an illustration). In the Computer condition, they made predictions about a computer vision system called "SOCAR" that was briefly described as a machine vision attached to a vacuum operated "lifting" system. A computer attached to a camera was shown as the illustration. Finally, in the Robot condition, subjects were shown a Robot named "OSCAR" that had a mechanical appearance, but an anthropomorphic shape with a head, arms, and legs. The wording of the scenarios was modified slightly to use either intentional language for the human and robot, and more passive mechanical language for the computer.

Subjects were given packets with color illustrations to read the scenarios from, and responded using a separate response sheet.

VIII. Experiment 2 Results

Subjects gave significantly more intentional responses for the human's behavior (59%) than the computer's behavior (38%), $t(19)=2.31$, $p=.032$. The robot predictions were intermediate between these (45%), but significantly different from neither. Although the robot mean across the 6 scenarios was between the other means, this was true for none of the individual scenarios. As Figure 1 makes clear, the proportion of intentional responses for the robot was either less than the proportion for the computer, or more than the proportion for the human.

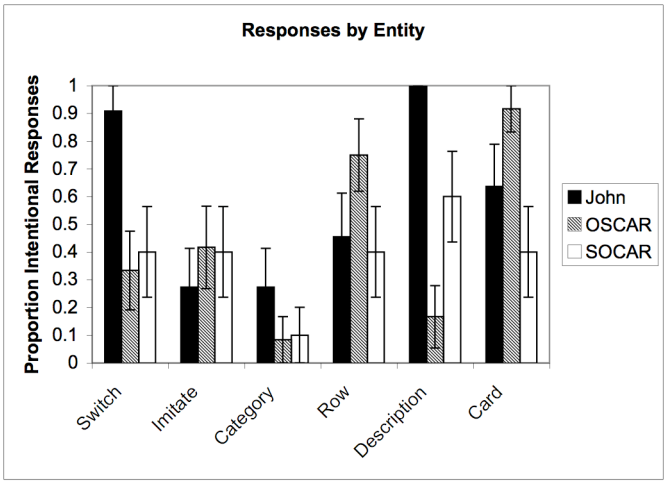

Figure 1. Proportion of intentional responses to each of six behavioral prediction scenarios.

IX. Experiment 2 Discussion

In this experiment, subjects gave different predictions for human and computer actors consistent with the hypothesis that they impute intentionally guided action for the human and nonintentional action for the computer. This provides support not only for the hypothesis that subjects use different concepts to understand human and computer driven behavior, but also that these concepts have an impact on the specific predictions they make about the behavior of these systems. The specific pattern of results across scenarios was relatively consistent for the comparison between the computer and human actors - in five of six cases, subjects made more intentional predictions for the human than the computer.

Although results contrasting the human and computer were consistent, reasoning about the robot was much more variable. In three of the scenarios the robot was seen as very similar to, or even less intentional than the computer, and in two of the cases it was seen as more intentional than the

human. This pattern of results suggests that subjects were engaging in ad hoc reasoning for the robot, making their decision about what it would do on a case-by-case basis instead of using more general knowledge about robots to guide decisions in a consistent way across scenarios. Research on concept formation does suggest that subjects sometimes create ad hoc categories in the service of the immediate goal of understanding a given situation [8]. Another option, however, is that subjects do have a more systematic explanation of the representational processes inherent to robots, but that it is organized around dimensions that vary orthogonally to the intentional/nonintentional contrast we have tested here. Future research might get at this issue either by testing individuals with more knowledge of robots to determine whether they are guided by a consistent domain theory, or by searching for different dimensions upon which behavioral predictions might vary for robots.

X. Conclusion

These studies suggest that adult reasoning about representational systems is shaped by a belief-desire folk psychology that has at its base a distinction between intentional and nonintentional representational systems. In Experiment 1, subjects not only differentiated intentional (human) and nonintentional (computer) representational systems, but they did so specifically for intentional mental states. In Experiment 2, subjects made fundamentally different predictions about the behavior of these systems, predicting more goal-directed behavior for the human relative to the computer. In addition, initial data on the robot, a system with a mixture of a features characteristic of intentional and nonintentional systems suggests that subjects do not simply choose an intermediate model for them, but rather tailor their responses to the specific situation. This pattern of responding may occur because, in contrast to people and computers, subjects lack a coherent, internally consistent model for the capabilities of robots, and instead reason on an ad hoc basis as new situations present themselves.

In addition to the applied implications of these findings, we would like to point out the more basic way in which they can inform not only research on the adult folk psychology, but also theory in cognitive development. First, these findings make a clear connection between early emerging understandings of representation, and the later adult understanding of the same issue. Early differentiation between human and mechanical representational systems can be used to understand how adults organize these categories later in life. Many researchers assume as much in arguing that understandings of early knowledge will provide a basis for organizing an understanding of adult knowledge by discovering the concepts that underlie, organize, and constrain adult thinking. However, this assumption is rarely tested, and in some cases adult understandings may not be the richly elaborated extension of early knowledge that is reflected in default assumptions about cognitive development. For example, based on research exploring early representational understanding, one might predict that adults would be able to effectively reason about visual attention, and understand that their mental representation or a scene is not a copy of it, but is a reduced and abstracted version of the scene. In stark contrast to this prediction, we have found that adults make very large mispredictions about their ability to perceive visual changes in a wide variety of circumstances, responding as if they had a copy of the visual world in their head [6].

However, in this context it is important to note that in a sense, we are turning the developmental data on its head by arguing that the concepts we have identified cause differential predictions about behavior because most theories of the development of TOM suggest that the behavioral predictions are a precursor to the full concepts. At this point, we have not revealed the causal connections between intentional/nonintentional explanatory frameworks, and specific behavioral predictions, so our hypotheses relating the two are necessarily speculative. It is certainly plausible that subjects reason from behavior to create ad hoc TOM-like explanations for all of the systems, not just the robot. However, one factor that makes this seem unlikely is that the induction scenarios in Experiment 1 were relatively divorced from specific behaviors making a behavior-first chain of reasoning awkward. One option might be to hypothesize a developmental sequence whereby initial learning is behavior-based, but with conceptual maturity, the concept-behavioral prediction link changes to allow a rich interchange between progressively more independent, but structurally similar, reasoning systems. Thus, in adults' conceptual induction and behavioral predictions allow both for abstract knowledge to inform an understanding of specific behaviors, and for a more behavior-based attributions about the essential nature of the internal processes driving the behavior. In other words adults can both take evidence about what something does to learn about how it thinks, and use understandings about thinking to predict what something does.

These findings show that adults classify representational systems and make behavioral predictions about them using concepts that vary in the degree to which they attribute intentionality to different kinds of agent. In doing so, they converge with other recent data showing that people segment action streams differently when they imagine a human or computer audience [9], and move differently and produce more social gestures when actually demonstrating actions for human and computer audiences [10]. In addition, future experiments along these lines may inform current work in robotics exploring how people perform perspective taking with robots [11], and how they attribute different kinds of knowledge to these systems [12]. Thus, this research may have implications both for human-machine interaction and cognitive development. In the former case, this research can serve as the basis for understanding people's expectations about the capabilities of the systems they interact with, and in the latter case it can help understand the developmental endpoint of well-understood early developing TOM.

References

[1] B. Reeves, and C. Nass, The Media Equation, New York: Cambridge University Press, 1996.

[2] A.L. Woodward, “Infants selectively encode the goal object of an actor’s reach,” Cognition, Vol 69, pp. 1-34, 1998.
[3] A.N. Meltzoff, “Understanding the intentions of others: Reenactment of intended acts by 18-month-old children,” Developmental Psychology, Vol 31, pp 838-850, 1995.
[4] V. Kuhlmeier, P. Bloom, & K. Wynn, “Do 5-month-old infants see humans as material objects,” Cognition, Vol 94, pp 95-103, 2005.
[5] E.S. Spelke, A.T. Phillips, and A.L. Woodward, “Infants’ knowledge of object motion and human action” in Causal Cognition: A Multidisciplinary debate, D. Sperber, D. Premack, & A Premack, Eds. Oxford University Press. 1995.
[6] D.T. Levin and M.R. Beck, “Thinking about seeing: Spanning the difference between metacognitive failure and success” in Thinking and Seeing: Visual Metacognition in Adults and Children, MIT Press, 2004.
[7] A.L. Woodward, “Infants selectively encode the goal of an actor’s reach.” Cognition, Vol 69, pp1-34, 1998.
[8] L.W. Barsalou, “Ad hoc categories,” Memory and Cognition, Vol 11, pp211-227, 1983.
[9] S. Killingsworth, M.M., Saylor, and D.T. Levin, “Segmenting action for computers and humans: Possible links to intentional understanding”. *Proceedings of the 14th Annnual IEEE International Workshop on Robot and Human Interactive Communication*, Vol 14, pp 196-201, 2005.
[10] J.S. Herberg, M.M. Saylor, D.T. Levin, P. Ratasnaswasd, and M. Wilkes, “Social and motor behavior in action demonstrations influenced by intentionality of audience”. Unpublished, 2005.
[11] G. Trafton, A. Schultz, M. Bugajska, and F. Mintz, “Perspective-taking with robots: Experiments and models.”*Proceedings of the 14th Annnual IEEE International Workshop on Robot and Human Interactive Communication*, Vol 14, 2005.
[12] S. Kiesler, “Fostering common ground in human-robot interactions.”*Proceedings of the 14th Annnual IEEE International Workshop on Robot and Human Interactive Communication*, Vol 14, 2005.

Appendix: Complete Set of Intentional and Nonintentional Mental Properties.

The following is a complete listing of the intentional and nonintentional mental properties used in Experiment 1. Note that each was presented as an induction from the person for some subjects, and from the computer for other subjects.

Intentional

Imagine that the Dell personal computer attempts to predict what a person is going to do next using an "M-rule" that takes their specific preferences into account using an algorithm built around a set of Gaussian kernels.

Imagine that John's brain uses a special rule to determine what specific kind of information another person wants. This rule is an IR-rule.

Imagine that the Dell personal computer attempts to determine whether a person does not know some important bit of information by completing a series of inferences about their behavior called a "Rhinnean set".

Imagine that John's brain attempts to determine when a person is working under a false assumption by using a stimulus-response procedure. This procedure requires that previous behaviors be correlated with known stimuli using hierarchical set analysis.

Nonintentional

Imagine that John's brain can keep a large amount of information for a long time, and sometimes organizes it by grouping things temporally (e.g. by putting things together that occurred at about the same time) using a regression-based principle.

Imagine that some processes within the Dell personal computer can only handle a few bits of information at a given time. This limit is referred to as a core striction.

Imagine that in some cases, the Dell personal computer encodes internal information using a "labeled line" code.

Imagine that John's brain sometimes transmits internal information more quickly at the expense of accuracy using a "delta rule".