A Method for Evaluating Users' Understanding of XAI Systems: The Mental Model Matrix

Gary Klein
Joseph Borders
Macrocognition, LLC

Robert R. Hoffman
Institute for Human and Machine Cognition

Shane Mueller
Michigan Technological University

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Cite as:
Klein, G., Borders, J., Hoffman, R.R. and Mueller, S.T. (2021). "A Method for Evaluating Users' Understanding of XAI Systems: The Mental Model Matrix. Technical Report, DARPA Explainable AI Program.

Introduction to the Problem

A Technical Report on "Measuring Mental Models" (Hoffman, Mueller, Klein and Litman, 2018) asserted that a methodology is needed for eliciting, representing, and analyzing users' mental models of intelligent systems., in part to enable researchers to tell whether the explanations adduced by their XAI system have been effective.

There is a large body of research in cognitive science on mental models (cf. Johnson-Laird, 1983; 1989), and it pertains directly to how people understand machines, spanning simple devices, process control systems, complex computational systems, and intelligent systems (for a recent review, see Hoffman, et al., 2018). Although the consensus view in the field of cognitive science is that mental models are an important abstraction for understanding reasoning, there is less agreement about how and whether they can be elicited or evaluated. For example, Nisbett and Wilson (1977) suggested that humans have little or no direct introspective access to higher mental processes. The most famous rejoinder to this work (Ericsson and Simon, 1980) is optimistic and discusses some methods of eliciting some aspects of mental models.

In applied cognitive science and human factors research, substantial methodological work has demonstrated the usefulness of a variety of methods for eliciting mental models (Crandall, Klein, and Hoffman, 2006), which involve a number of methods such as think-aloud problem solving, retrospection, diagramming, and other procedures as well. These are all forms of "self-explanation" in that they engage the user in expressing their own understanding. Methods for representing user mental models range from free text to causal diagrams to formal (propositional) representations of knowledge and beliefs. Essentially, all of these methods derive from what we can call the "Correctness Assumption"— *that a representation of a user's mental model of a complex system is a knowledge structure that correctly describes how they believe system works.*

We all have mental models for different types of systems, machines, and organizations and even for social interactions. Our mental models provide us with a blueprint for how the device or the interaction produces its results. They let us describe a system's form, explain how it functions, and predict its future states (Rouse and Morris, 1986). However, mental models are limited in that they are reductive (simplifications) and typically are incomplete. Furthermore, people sometimes overestimate how well they understand complex causal systems (Chi, et al, 1989). Knowledge about complex systems is often piecemeal (diSessa, 2018). They can also be resistant to change (Mueller and Tan, 2018), even in light of disconfirming evidence. For example, "Knowledge shields" are arguments that learners make that enable them to preserve their reductive understandings (Feltovich, Coulson, and Spiro, 2001). A focus for instructional design has been to develop methods to get people to recognize when they are employing a knowledge shield that prevents them from developing richer mental models.

For effective use in the development of XAI systems, and successful training for end-users, what is most desirable is a mental model elicitation task and accompanying representational scheme that enables the researcher to see what is good *and also* what is incomplete or incorrect about a user's mental model, and enable the learner to learn what is good *and also* what is limited in their understanding.

Finding a Solution to the Problem

Two of the elicitation methods presented in Hoffman, et al. (2018) (see Table 4) were directly suggestive of a need to re-examine the Correctness Assumption: the Glitch Detector Task and the ShadowBox Task. These methods focus less on revealing the user's understanding of "how the system works" and more on how the user's mental model is itself limited and how the system *does not* work. Users' mental models often do capture system limitations and failings, as operators often report a sensitivity to system weaknesses and brittle points. How can XAI system developers and trainers help users discover and explain aspects of their mental model that are reductive or incorrect?

In the Glitch Detector Task (Hoffman, et al., 2001; Taylor, 1988), users identify the things that are wrong with their understanding or wrong with an explanation. Similarly, in the ShadowBox Task (Klein and Borders, 2016), users compare their understandings and explanations to those of a domain expert. In the method, the user is presented a question such as "How does a car's cruise control work?" Accompanying the question is a proposed explanation (derived from explanations provided by an expert). The task for the user is to identify one or more ways in which the explanation is good, and ways in which it is bad. After doing this, the participant is shown a Good-Bad list that was created a domain expert. The participant's comparison of the lists can lead to insights. Table 1 (below) presents an example result from an application of the ShadowBox Task.

Table 1. An example of propositional coding using the ShadowBox Task.

The control unit detects the rotation of the drive shaft from a magnet mounted on the drive shaft, and from that can calculate how fast the car is going. The control unit controls an electric motor that is connected to the accelerator linkage. The cruise control adjusts the engine speed until it is disengaged.	
What is right and helpful about this explanation?	*What is problematic or wrong about this explanation?*
The cruise control unit has to know how fast the car is going. The cruise control has to control the engine throttle or accelerator.	It seems overly technical, with some concepts left unexplained. I do not think the cruise control detects the engine speed.

Borders, Klein and Besuijen (2019) observed and interviewed industrial process control operators as they responded to upset scenarios on a high fidelity training simulator. Some of the operators had more than a decade of experience, but most had less than three years and one had only six months; they averaged 4.5 experience. The scenarios were very demanding. And no two operators approached them in the same way.

How the System Works

As expected, we found that the operators relied on a set of beliefs about how the system worked. Sometimes these beliefs were limited in ways the operators didn’t appreciate, and sometimes they

were flawed, but generally they were accurate and the operators usually were able to diagnose their own confusions.

How the System Fails

We also found that operators understood ways that the system could fail — its limitations and its vulnerabilities to breakdown (see also Mumaw, et al., 2000). These "negative" beliefs were a very important aspect of the operators' mental models — providing them with ideas about what might be going wrong. Being able to consider and anticipate system limitations and failures is obviously important for troubleshooting. Imagining how a system might fail rather than just considering how it is supposed to work is also a very important aspect of system design. Too many designers fixate on delivering a system that meets the requirements, and don't stop to imagine where the system might break down, the conditions under which a system might crash. Mumaw, et al. (2000) found that workers monitoring a nuclear power plant couldn't just rely on the schematics. They had to appraise the plant's performance against a noisy background. They had to be alert to recent developments such as valves that were sticking or sensors that were acting up.

Workarounds. The operators had beliefs about how to do workarounds to overcome limitations and failures. These workarounds were important for recovering from upsets. Knowing how to perform workarounds is obviously important for adapting to unexpected situations. The more experience operators had, the more sophisticated were their ideas for keeping the system running.

Confusions. Finally, a representation of a user's mental model should include beliefs about the limitations of people, such as the users of a system — the ways they can become confused. For example, someone might direct us to a location (e.g., Go two blocks, turn left, etc.), but a person with a stronger mental model of the route and of our navigation abilities might anticipate where we might get confused or mistaken and annotate the directions accordingly (e.g., Go two blocks and turn left; it's a narrow street and there's no street sign so it might look like a driveway, but there's a little antique store on the far corner). Here, the mental model is about our limitations and potential failures, not those of a system. It can be quite impressive when people can anticipate the ways that others might get confused, and make the appropriate adjustments.

These are kinds of findings were suggestive of a concept we now refer to as the "Mental Model Matrix" (MMM; Klein, 2021) — which emphasizes the important aspects of mental models that often get ignored.

The Mental Models Matrix

Figure 1 presents the Mental Model Matrix.

	Positive	Negative
AI/XAI System	(1) How the System works: Parts, connections, functions, relationships, control logic	(3) How the system Fails: Breakdowns, limitations
The User/Learner	(2) How to make the System work: Detecting anomalies, appreciating the System's responsiveness, performing workarounds and adaptations	(4) How the User/Learner gets confused: The kinds of errors the User made, or other Users might make

Figure 1. The Mental Models Matrix.

The initial concept of a mental model — a set of beliefs about how a system works — is certainly valuable. But if it is guided only by the Correctness Assumption, it is incomplete. Indeed, it misses the crucial kinds of beliefs, gained through experience, that underpin an expert's kill.

Example of a Completed MMM

We referred above to the study by Borders, Klein and Besuijen (2019), which was a first use of the MMM concept. The researchers observed eight highly qualified process control operators independently complete two challenging and unfamiliar scenario exercises on a high-fidelity training simulator. After they completed each scenario, they participated in cognitive interviews in attempt to formalize their mental model with respect to a complex manufacturing process and the associated control panel. As part of the interview they were asked about their approach to managing the upset scenario, including gaps in their understanding, and limitations/strengths to their diagnoses and decision making.

Through the observations and interviews, a Mental Model Matrix was created that spanned their knowledge of how the system functions (+ system), including the parts and connections among those parts that make it work. The MMM also included the operators' working knowledge for the limitations of the system (- system), which is the byproduct of their extensive experience working within the system and confronting its boundary conditions and edge cases.

In addition to their knowledge of the system, the operators were intimately familiar with how they could control and manipulate how it worked (+ human). The operators used strategies for detecting

and responding to anomalies, and they used the affordances available to them to maintain control of the system. Finally, some operators were also aware of the kinds of difficulties to which they (and other operators) were susceptible. As a result, their mental model included knowledge for their tendencies and vulnerabilities that put them and the process at risk, such as inert knowledge, fixation, and explaining symptoms away. Figure 2 shows the MMM that integrates the material from the eight participating operators.

	+	-

System	• Fluidics principles
	  ○ Pressure differentials
	• Thermodynamics
	• Vessels & subsystems
	  ○ Distillation column (splitter)
	  ○ Reflux drum
	  ○ Exchangers
	  ○ Reboilers (primary, secondary)
	  ○ Feed (input) & product draws (output)
	• Analyzers/transmitters
	  ○ Flow rates (input/output)
	  ○ Pressure differentials
	  ○ Temperatures
	  ○ Levels
	  ○ Bottoms draw
	• Advanced control logic (e.g. process deviation monitors)
	• Trends
	• Alarms
		• Inaccurate data caused by drifting and/or failing instrumentation (e.g. transmitter and analyzers)
		  ○ Temperature, flow, pressure, level instrumentation is sensitive to weather and general wear and tear
		• Valve control malfunctions
		  ○ Loss/limitations in peripheral systems (e.g. air supply)
		  ○ Operator errors (field operators not setting valves to the correct placement)
		  ○ Blockages/build-up
		• Faulty advanced automated systems and control logic
		  ○ Automation is not always sensitive to contextual factors, global system states, and faulty instrumentation

Human / Operator	• Problem detection strategies
	  ○ Knowledge of normal operations provides a baseline for spotting deviations
	  ○ Managing alarms
	  ○ Verify data to ensure sensors are accurate (using secondary sensors)
	  ○ Probing/testing (e.g. making small changes to see how the system reacts)
	• Diagnostic strategies
	  ○ Mass balance
	  ○ Probing/testing
	• Disturbance management
	  ○ Prioritizing disturbances
	  ○ Triaging response to buy more time in diagnosing the problem
	• Resource management
	  ○ Utilizing system features (e.g. secondary reboiler, flares, overfill vessels)
	  ○ Teamwork: getting new perspectives, triage tasks, seek information from the field
		• Inert knowledge
		  ○ Operators sometimes forget what they've learned, especially during challenging and time sensitive situations (procedures and documentation can be useful here)
		  ○ Operators can also fail to use vital resources when they become overloaded
		• Workload limitations
		  ○ Operators have a sense of the reciprocal relationship between their available resources and the number of resources demanded by managing the process
		  ○ Understanding when resource demands surpass one's available resource supply (i.e. red line) leads to performance decrements
		• Fixation
		  ○ The tendency to form an initial idea of what is going wrong and then hold onto this idea without testing it and despite contrary evidence
		• Distractions
		  ○ Being aware of and mitigating potential distractions to maintain high performance

Figure 2. A completed Mental Models Matrix, for the case of industrial process control.

Uses of the Mental Models Matrix

Although the MMM was originally conceived of as a descriptive tool, it can be used for system development, evaluation and training.

Tool for System Users

Users can be presented the matrix while they are engaged in a knowledge elicitation task. They could be presented with the MMM after having received initial instruction in the use of an XAI system, with encouragement to ponder aspects of their mental model that may be deficient. Thus, the MMM would be a tool to support elaborative Self-Explanation, letting the quadrants guide the user to enrich their mental models.

Tool for Designers

The MMM can allow designers think more broadly about their systems, and help them think more broadly about how system users may be (mis) understanding and adapting to the system. Too often, system designers content themselves with demonstrating that their product can work as advertised, without considering its boundary conditions, without considering whether they have provided any means for users to work around these boundary conditions, and without trying to imagine the conditions that might confuse the users. The MMM could be an important corrective to this type of design myopia. The MMM can provide System Designers with insights into the understanding and the misunderstanding of users. In other words, it may be possible to show designers and developers the goodness of the mental models of their intended Users.

Evaluative Tool

Two evaluative comparison tasks are possible.

(1). *The Developer-User Comparison Task.* Working independently, a set of researchers would contribute statements/propositions that would fit into each of the four cells of the MMM. A group of Users would do likewise. For both groups, the method they would use for elicitation might be unprobed Think-Aloud Problem Solving or Task Reflection (see Hoffman, et al., 2018). Statements expressing the User's mental model are then compared to statements provided by the Developers. This comparison can be done by the Developers, but it can also be done by the Users themselves. What is important in this latter form of the Developer-User Comparison Task is that some of the propositions in the Developer's MMM would be altered so as to be incorrect or inaccurate. The inaccurate propositions could serve as foils, designed to surface potential misunderstandings that system Users might have. (The preparation of these foils might itself be a constructive exercise for the system Developers.)

(2). *The Expert-User Comparison Task*. A small panel of domain experts would respond to the propositions in both Users' MMMs and the Developers' MMMs, indicating which of the propositions they agreed with and those which they found problematic. Then the evaluators could collect data from a target audience of representative users, and determine the general accuracy of the users' mental models as well as diagnosing the kinds of misunderstandings that users are showing. Some of the propositions in the Developer's or the Experts' matrix might have no analog in the user's matrix. It would be possible to generate frequency counts of such matches and

mismatches, and using those apply scaling and statistical analyses. In all of these evaluative methods, the completed matrices can be scored for correctness, and scaled on such factors as completeness, sophistication, etc. A small panel of Experts might provide evaluations of the goodness of Users' mental models (see Crispen and Hoffman, 2016).

Training Tool

The MMMs created to represent the mental models of Developers and of Experts would be ideal materials to use in a ShadowBox training exercise (described above). After users indicate which of the propositions they agreed with and which they found problematic, they could be shown the Experts' responses.

References

Borders, J., Klein, G., and Besuijen, R. (2019). An operational account of mental models: A pilot study. Presentation as the 15th International Conference on Naturalistic Decision Making, San Francisco, CA.

Crandall, B., Klein, G., Klein, G. A., Hoffman, R. R., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. MIT Press.

Crispen, P., and Hoffman, R. R. (2016, November/December). How many experts? *IEEE Intelligent Systems*, pp. 57-62.

diSessa, A.A. (2018). A friendly introduction to "Knowledge in Pieces": Modeling types of knowledge and their roles in learning. In G. Keiser, et al. (Eds)., *Invited Lectures from the 13th International Congress on mathematical Education* (pp. 65-84). Cham, Switzerland: Springer.

Ericsson, K. A., and Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215.

Hoffman, R.R., Coffey, J.W., Ford, K.M. and Carnot, M.J. (2001, October) STORM-LK: A human-centered knowledge model for weather forecasting. In J.M. Flach (Ed.), *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (p. 752). Santa Monica, CA: Human Factors and Ergonomics Society.

Feltovich, P. J., Coulson, R. L., and Spiro, R. J. (2001). Learners'(mis) understanding of important and difficult concepts: A challenge to smart machines in education. In K.D. Forbus and P.JK. Feltovich (Eds.), *Smart machines in education* (pp. 349–375). Menlo Park, CA: AAAI/MIT Press.

Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects." Technical Report from Task Area-2, DARPA Explainable AI Program, DARPA, Alexandria, VA.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Johnson-Laird, P. N. (1989). *Mental models.* In M. I. Posner (Ed.), *Foundations of cognitive science* (p. 469–499). The MIT Press.

Klein, G. (2021). The mental model matrix. Important aspects of mental models often get ignored. [blog]. *Seeing what others don't. https://www.psychologytoday.com/us/blog/seeing-what-others-dont/202101/the-mental-model-matrix*

Klein, G., and Borders, J. (2016). The ShadowBox approach to cognitive skills training: An empirical evaluation. *Journal of Cognitive Engineering and Decision Making, 10,* 268-280.

Mueller, S. T., and Tan, Y. Y. S. (2018). Cognitive perspectives on opinion dynamics: the role of knowledge in consensus formation, opinion divergence, and group polarization. *Journal of Computational Social Science*, *1*(1), 15-48.

Mumaw, R.J., Roth, E.M., Vicente, K.J., and Burns, C.M. (2000). There is more to monitoring a nuclear power plant than meets the eye. *Human Factors, 42*(1), 36-55.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Rouse, W.B., and Morris, N.M. (1981). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin, 100*(3), 349.

Taylor, J.R. (1988). Using cognitive models to make plants safer: Experimental and practical approaches. In I.P. Goodstein, H.B. Andersen, and S.E. Olsen, (Eds.) *Tasks, errors and mental models* (pp. 233-239). New York: Taylor and Francis.