An Abductive Approach to Covert Interventions

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)
School of Biomedical Informatics, University of Texas Health Science Center at Houston
7000 Fannin Suite 600, Houston, TX 77030, USA

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)
School of Biomedical Informatics, University of Texas Health Science Center at Houston
7000 Fannin Suite 600, Houston, TX 77030, USA

Abstract

We explore ways of covertly delivering interventions into the adversary decision cycles so as to effectively shape adversary decision-making and performance without inducing much suspicion. Recognizing that completely covert interventions, while most effective, are difficult to implement, we focus on a more general mode of covertness. Based on insights from human abductive reasoning, we propose a delivery scheme that contains interventions that may be noticeable but whose true meanings are hidden or distorted (e.g., the human operators do not easily attribute the interventions to malicious attacks). We evaluate, both theoretically and empirically, the effectiveness and robustness of this scheme in escaping detection and disrupting performance.

Keywords: Abduction, decision making, cybersecurity, intervention.

General Formatting Instructions

A cyber attack is more damaging and harmful if it is stealthy and escaping detection. One critical challenge in cyberspace security is therefore to find ways to effectively detect hidden or covert attacks. One approach to meeting the challenge is to look at the other side of the coin and study how and why some attacks can be delivered covertly that induce no or minimal suspicion from the human operator. The results from this approach can then be used to design better countermeasures and improve security.

Here we focus on the concept of "covertness" in cyber attacks and intend to discover the theoretical essence and practical guidelines of implementing "covertness". Presumably, a covert attack would be one that is completely hidden and not noticed by the targeted operator at all. In this sense, covertness can be implemented as slip of attention. Examples include attention blink, change blindness, and inhibition of return, to name a few. While a large body of evidence has confirmed that attention is a fragile cognitive function that can be manipulated and exploited for the purpose of implementing covertness, it has also been suggested that the attention-based approach is quite limited and difficult to apply in the real world situations.

There is at least another mode of covertness. In this mode, signs of the intervention are noticed by the human operator (therefore, the intervention is not completely hidden and escaping attention), however, the true meaning/significance of the intervention is disguised or distorted or hidden in such a way that they do not easily result in suspicion of outside influence. This mode of covertness suggests new ways of implementing covertness.

Consider the following scenario: It is 12am and that John, an analyst, is working on a sensitive document on his computer and you have delivered a virus to his computer in order to take a peek. Ideally, you would like your operation is completely invisible to John, but unfortunately, one inevitable side effect of your virus is that John's computer becomes slow, which *John eventually notices and starts to become suspicious.* Then John receives an alerting pop-out message informing him that the antivirus software on his computer has started scanning as scheduled and that so far no virus has been found. John now understands why his computer becomes slow, is relieved, and continues to work on his document, without realizing your peeking eyes.

Though hypothetical, this example highlights an important aspect of covertness, which has to do with an understanding of how a human operator reasons and explains unexpected observations and if and when the operator becomes suspicious given data. In the example, John becomes suspicious when he notices that his computer slows down, an often-inevitable indicator of attacks. But his suspicion fades away after the pop-out message, which is also delivered by the attacker with the goal to provide a better explanation for the slow-down so as to reduce John's suspicion.

Instead of directly exploiting the low-level attentional function, this mode of covertness depends upon exploiting a higher-level human inference system called abduction. We argue that this mode of covertness is more general, more realistic and potentially more powerful.

Abduction-based Covertness

Abduction was introduced by American philosopher Charles S. Peirce (1839-1914) as a form of human inference that is different from deduction and induction. According to Peirce, in abduction one infers causes from effects or explanations from observations (See Fann, 1970 for a general introduction to Peirce's theory of abduction). The general form of abduction is shown below,

A fact C is observed,
H can explain C;
Hence, H may be true.

Here is a specific example of abductive inference, in the context of the hypothetical scenario above,

The computer suddenly slows down,
A malicious attack explains the slowdown;
Hence, a malicious attack may be occurring.

And here is another example,

The computer suddenly slows down,
Antivirus scanning explains the slowdown;
Hence, nothing is wrong and just be patient.

Charniak and McDermott (1985) characterize abduction as modus ponens turned backward (see also Brachman & Levesque, 2004). It is clear in abduction the conclusion does not necessarily follow the premises – in the above examples two different explanations are inferred to explain a same observation. However, according to Peirce, abduction is important in that it "is the only logical operation which introduces any new ideas; for induction does nothing but determine a value [to classify], and deduction merely evolves the necessary consequences of a pure hypothesis" (Peirce, 1931, v. 5, p. 171). Though inconclusive, the explanation inferred by abduction "is adopted for some reason, good or bad, and that reason, in being regarded as such, is regarded as lending the hypothesis some plausibility" (Peirce, 1931, v. 2, p. 511).

Modern researchers often regard abduction as a complex process of finding a best explanation for a set of observations (Josephson & Josephson, 1994; Paul, 1993; Thagard, 1992). Since "explaining" is such an inevitable aspect of human everyday activities, abductive reasoning is almost ubiquitous, ranging from hearing the thunder ("It's going to rain?"), seeing a falling maple leaf ("Autumn has come?"), to medical diagnosis (from symptoms to diseases) and scientific discovery (from data to knowledge and theories). In battlefields, commanders have to infer the enemy's motivations based on observations and intelligence and then take proper actions. In cyberspace security, operators have to infer if an attack has occurred given observations.

How do people do abduction? The Theory of Explanatory Coherence (TEC) is an influential theory of human abduction (Thagard, 1989, 1992). According to TEC, abduction is a parallel constraint satisfaction process in that all propositions, including explanations, evidence, and explanatory relations, form a network that constantly seeks harmony. An explanation should be accepted if it is coherent with all other propositions in the network, rejected if it is incoherent, and the best explanation for available observations is the one that enjoys the most explanatory coherence in the network. TEC proposes seven principles that establish explanatory relations among propositions and regulate the global coherence of an explanatory system: (1) *symmetry*: If P and Q cohere, then Q and P cohere; If P and Q incohere, then Q and P incohere. (2) *explanation*: If $P_1 \ldots P_m$ explain Q, then $P_1 \ldots P_m$ cohere with each other and with Q cohere, and the degree of coherence is inversely proportional to the number of propositions $P_1 \ldots P_m$. (3) *Analogy*: If P_1 explains Q_1, P_2 explains Q_2, P_1 is analogous to P_2, and Q_1 is analogous to Q_2, then P_1 and P_2 cohere, and Q_1 and Q_2 cohere. (4) *data priority*: Observations have a degree of acceptability of their own. (5) *Contradiction*: If P contradicts Q, then P and Q incohere. (6) *competition:* If P and Q both explain a proposition, and if P and Q are not explanatorily connected, then P and Q incohere. (7) *acceptability:* The acceptability of a proposition P depends on its coherence with all the propositions in the system.

TEC has been computationally implemented in a connectionist system called Echo (Thagard, 1992). In Echo, propositions (both data and hypotheses) are represented by nodes. Coherence relations are represented by excitatory links and incoherence relations are represented by inhibitory links. Node activation represents the node's degree of coherence with all propositions in the network. The system updates itself based on parallel constraint satisfaction (Thagard, 1992). During this process, propositions that are incoherent die out and propositions that are coherent are strengthened. In the end, the most activated propositions represent the most plausible and coherent explanations. Echo has been extended to UEcho to incorporate more sophisticated handling of uncertainty (Wang, Johnson, & Zhang, 1998; Wang, Johnson, & Zhang, 2006).

TEC and UEcho capture several critical constraints in abduction, including explanatory breadth (the model prefers a hypothesis that explains more); simplicity (the model prefers a simpler hypothesis); being explained (the model prefers a hypothesis which itself is explained); data reliability (the credibility of an observation also depends on its coherence in the system); and analogy (analogous hypotheses are coherent). More important, however, they shed interesting new insights on human suspicion and implementing covertness. In this context, suspicion can be viewed as the degree of acceptance of an explanation such as "a malicious attack is occurring", and implementing covertness is not much more than to make the degree of acceptance of this explanation as low as possible.

Based on this reasoning, we hypothesize that effective covert interventions can be delivered in such a way that suspicion-bearing explanations (e.g., "a malicious attack is occurring") cannot become the best (winning) explanation given data. TEC has already offered several straightforward ways to do just this. For example, one way is to "explain away", which says that when delivering an intervention, deliver an explanation for that intervention as well so that the true meaning of observations can be shielded. This is exactly what happens in our previous hypothetical example. an attack is delivered, which caused John's computer to be slower. In anticipating this, a secondary message is delivered to John to "explain" to him that why his computer became slower. This new explanation "explained away" the John's observation and therefore reduced his suspicion – that is, the acceptance of "an attack is occurring" became low. Another example of abduction-based covertness derives directly from the data reliability principle – we can discredit those "suspicion-inducing" observations by introducing conflicting data ("unreliable data"). "Are you sure that your computer becomes slower?" By inducing new

data to promote John to cast his doubt, the suspicion level of "an attack is occurring" is reduced.

To a certain extent the attention-based covertness (i.e., delivering interventions that are invisible to the human attention) is a special case of this new abduction-based covertness. Since abduction starts with observations (that is, the data to be explained, e.g., "the fact C is observed"), completely invisible interventions suggest that suspicion-generating abduction will not even be starting in the first place. However, abduction-based covertness is more general in the sense that in case some suspicion-inducing observations become available, covertness can still be achieved if proper measures are taken so that the suspicion-bearing explanation will not become the most plausible one.

Stealth and Disruption with IMPs

As a preliminary step toward evaluating the effectiveness of abduction-based covertness, we conducted a study to examine how a human operator digests unexpected interventions and adjusts his level of suspicion. The study utilized a so-called Interface Manipulation Protocol (IMP). The toolbox of IMP contained dozens of possible intervention types that could be delivered to the adversary computers to cause disruption with, for example, keyboard and mouse operations. We were interested in finding out the optimal chain of IMP delivery scheme (e.g., when to deliver what IMP for how long?) that causes maximal disruption with minimal suspicion.

Method

Participants

Nine college students and graduate students in the Houston medical center area were paid to participate in the experiment.

Procedure

The experiment was programmed in E-Prime and conducted on a computer with a 20 inch LCD monitor. Subjects were instructed to type in sequences of random numbers as prompted (Figure 1). Table 1 shows three independent variables manipulated in the study, including the type of IMPs and the type of delivery themes.

Figure 1. Subjects were required to reproduce the target sequence. Errors were prompted in red color and need to be corrected with extra keystrokes. Errors could include "IMP errors" (produced deliberately by IMPs) and "genuine errors" (subjects' own typos). In this example, the subject had mistyped the target character number "9" with number "7", and subsequently typed number "8" before realizing the error.

At the beginning of each trial, subjects were first prompted with a sequence of 20 characters of random numbers, shown at the top of the computer screen. Then, they were to copy the entire sequence in exactly the same order, and their responses were shown one by one for each keystroke, in a separate line below the target sequence. If an error occurred, either by subjects' own error ("genuine errors") or by deliberate IMP interventions, the mismatched character would be shown in red. Subjects were instructed to immediately erase the error by using the backspace key. If subjects have skipped the error for several keystrokes, they had to erase all subsequently typed characters (including the correct ones). That is, correcting an error had to be done in a backward sequential order (similar to the situation of typing documents without the ability of adjusting the cursor position by mouse).

Table 1. Independent variables manipulated

3 types of IMPs • Non-responsive key (IMP1): when a key is pressed, nothing shows up, so the subject has to retype to correct; • repetitive key (IMP2): when a key is pressed, the same character will show up twice. For example, typing "3" would result in "33" shown on the screen, so that the subject has to erase the extra character; • altered key (IMP3): when a key is pressed, a randomly selected different character is shown (e.g., typing "3" and "4" shows up), so the subject has to erase the wrong character and retype.
4 types of delivery themes • Pure: only one type of IMPs is delivered. • Mixed: multiple types of IMPs are delivered. • Clumped: IMPs are delivered consecutively. • Dispersed: IMPs are delivered sparsely.
4 experimental conditions (combination of themes) • PC: Pure-Clumped. • PD: Pure-Dispersed. • MC: Mixed-Clumped • MD: Mixed-Dispersed
4 levels of IMP delivery rates 10%, 20%, 30%, 40% of the target characters are affected by IMPs).

Three types of IMP interventions were silently delivered by hijacking the subject's keyboard (see Table 1 and Figure 2). Each delivery of IMP intervention was designed to affect only one keystroke. For instance, by IMP1 (non-responsive key, Figure 2A), when the subjects typed any key on the keyboard (not necessarily matching the target character), the program would silently remove the keystroke such that no response character would be shown below the target character. Then, the IMP intervention would be temporarily

"disarmed" for this particular target character. Only when subjects pressed the same key for second time, the character would show up in the response line.

Four delivery themes were designed based on the mixture of IMP types and the temporal intervals between each IMP intervention (Table 1). In the "Pure" theme, only one IMP type was implemented for the target sequence. In the "Mixed" theme, all three types of IMPs were implemented. In the "Clumped" theme, IMPs were clustered together such that one IMP intervention could be immediately followed by another. In the "Dispersed" theme, IMPs were evenly distributed among the 20 target characters.

The delivery themes were grouped into 4 experimental conditions with each condition containing one particular combination of the mixture and temporal distribution (Table 1). For example, in the "PC" condition, only one type of IMP was delivered but in a clustered fashion. We also implemented 4 levels of IMP delivery rates, which were evenly distributed in each of the delivery themes.

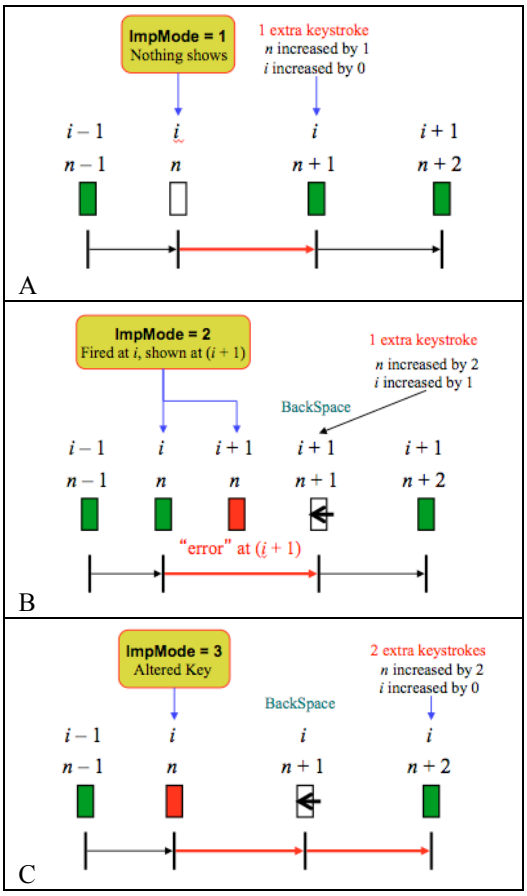

Figure 2. How types of IMPs (A: IMP1; B: IMP2; C: IMP3) affect dependent measures. *i* = position in the target sequence; n = number of keystrokes.

After practice trials, each subject completed 4 blocks of trials, with each block corresponding to one of the experimental conditions. The order of the conditions was randomized between subjects. Each block consisted of 20 trials, and each trial consisted of 2 target sequences. At the end of each block, subjects were asked to evaluate the "reliability" of the input device on an 1-to-7 scale with "1" for the most "unreliable" and "7" for the most "reliable". Subjects took a brief break before moving to the next block of trials.

There were two major dependent measures. Stealth (covertness) was measured by the subjective evaluations of the reliability of the input device. Higher evaluation scores indicated higher tolerance of IMPs and therefore less suspicion. Disruption was measured by the number of extra keystrokes ("ExtraKS") required to complete the sequence (excluding the extra keystrokes directly caused by IMPs). Higher scores of ExtraKS indicated more severe disruptions to the performance. The relation between IMPs and dependent measures is depicted in Figure 2.

Results

One main result of the study is shown in Figure 3, which depicts the effect of delivery themes on stealth and disruption. Statistics show that in terms of stealth the mixed-clumped delivery (IMPs with mixed types are delivered continuously) is the best (mean evaluation scores = **4.36**, 3.94, 3.76, 4.01, with standard errors = 0.22, 0.31, 0.19, 0.25, for MC, MD, PC and PD, respectively). And in terms of disruption the pure-dispersed delivery (IMPs with the same type are delivered sparsely) is the best (mean disruption scores = 2.06, 3.08, 2.62, **3.50**, with standard errors = 0.59, 0.58, 0.47, 0.73, for MC, MD, PC and PD, respectively). Further analysis shows that if we combine the two dependent measures, the pure-dispersed delivery has the highest effectiveness score, as shown in Figure 4 (mean effectiveness scores = 0.43, 0.51, 0.49, **0.56**, with standard errors = 0.03, 0.02, 0.03, 0.02, for MC, MD, PC and PD, respectively). Overall, it seems that the pure-dispersed delivery is the most effective IMP delivery theme if the tradeoff between stealth and disruption is considered.

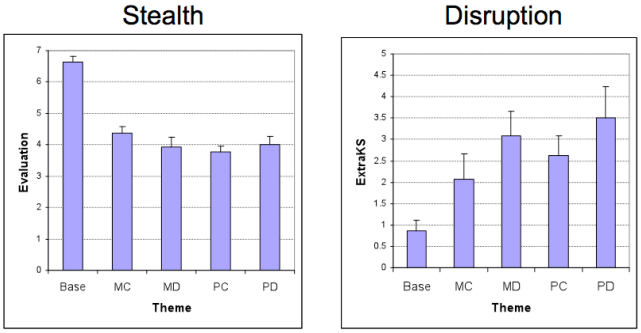

Figure 3. The effect of delivery themes (Base: no IMP was delivered; MC: IMPs were delivered in mixed-clumped fashion; MD: mixed-dispersed; PC: pure-clumped; PD: pure-dispersed).

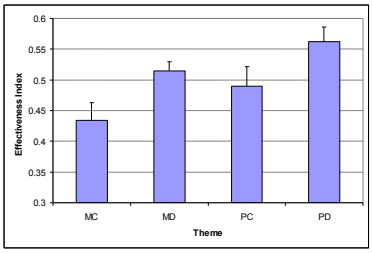

Figure 4. Effectiveness by themes. The effectiveness index is computed by adding the normalized stealth and disruption measures.

The effect of IMP types is shown in Figure 5. In terms of stealth, statistics show that IMP1>IMP2 (mean evaluation scores = 4.34, 3.70, with standard errors = 0.21, 0.23, for IMP1 and IMP2, respectively, $p<0.05$) and IMP1>IMP3 (mean evaluation scores = 4.34, 3.62, with standard errors = 0.21, 0.31, for IMP1 and IMP3, respectively, $p<0.05$). In terms of disruption, no significant difference is found (mean disruption scores = 2.63, 3.21, 3.34, with standard errors = 0.44, 0.68, 0.71, for IMP1, IMP2 and IMP3, respectively).

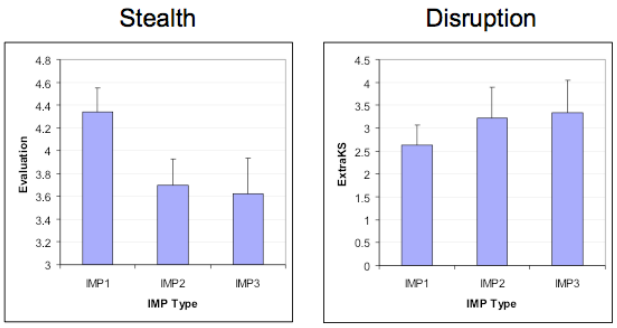

Figure 5. Effect of IMP types on stealth and disruption.

The effect of IMP delivery rate is shown in Figure 6. It is clear that with the rate increase the evaluation scores decrease (mean evaluation scores = 6.64, 5.12, 4.28, 3.58, 3.10, with standard errors = 0.17, 0.15, 0.23, 0.25, 0.28, for 0, 10%, 20%, 30% and 40%, respectively) and the disruption scores increases (mean disruption scores = 0.87, 1.77, 2.75, 2.97, 3.76, with standard errors = 0.25, 0.37, 0.50, 0.81, 0.59, for 0, 10%, 20%, 30% and 40%, respectively). A nonlinear regression supports the notion that rate increases led to more disruption and less stealth.

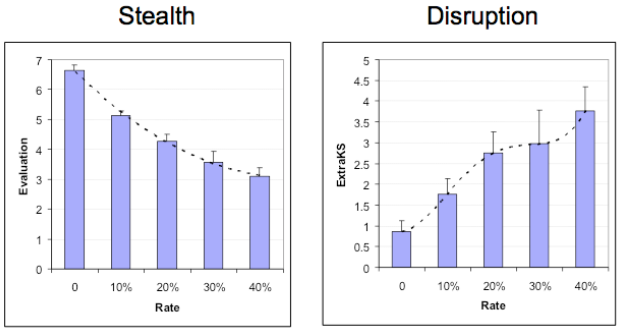

Figure 6. Effect of IMP delivery rates on stealth and disruption, with polynomial fitting curves.

Summary and Discussion

In this article we explore ways of covertly delivering interventions into the adversary decision cycles so as to effectively shape adversary decision-making and performance without inducing much suspicion. The focus here is not on the delivery technology, which we assume can be achieved, but on the covertness. That is, how can we deliver interventions that do not induce significant suspicion and effectively shape operators' behavior?

Attention is often the first cognitive faculty explored in the attempt to understand covert interventions. On the one hand, there are hardly better ways to implement covertness than designing interventions that are invisible even to the adversary operator's attentional system. In this case, the interventions are completely hidden and therefore can potentially cause most and long-term damage. On the other hand, a large body of evidence in the field of psychology has shown that attention is a fragile function that is subject to exploitation and manipulation. A recent theoretical breakthrough of attention research is the notion that there exist different types of attention, each of which is subserved by different brain regions and is sensitive to different variables (Fan, McCandliss, Sommer, Raz, & Posner, 2002; Posner, 2004). Equipped with the taxonomy, it has been suggested that each type of attention could be subject to different exploitations for the purpose of covertness. Studies have been conducted to systematically examine the effect of parameter changes on inducing covertness and affecting performance (Sun, Wang, Zhang, & Smith, 2008; Wang & Fan, 2007; Wang, Liu, & Fan, 2012).

Recognizing the limitation of attention-based covertness in real world situations, in this article we propose to a more general approach to covertness. That is, instead of delivering completely hidden interventions, it is possible to deliver interventions that may be noticeable but whose true meanings are hidden or distorted. Consequentially the similar effect of covertness can be achieved. This approach, based on insights from human abductive reasoning rather than straightforward attentional manipulations, is easier to implement and potentially more powerful. However, the success of the approach would require a better understanding of adversary decision processes and more sophisticated delivery strategies. The study reported in this

article is a step towards developing and evaluating guidelines and schemes for such deliveries.

In the experiment we manipulated the type of interventions and the delivery themes. In particular, we distinguished pure vs mixed and clumped vs dispersed deliveries. We evaluated the effect of these manipulations on suspicion and performance. Our results support the general notion of abduction-based covertness. We show that covertness can be achieved even when interventions are detected as long as they are not properly explained. Our results demonstrate that different intervention types have different effectiveness. And more important, we show that pure-dispersed delivery scheme is more effective than the other delivery schemes, suggesting that when delivering interventions, to achieve effective stealth and disruption, try to keep the interventions dispersed and do not mix different types of interventions.

In sum, we demonstrate that abduction is a sound and insightful framework for understanding human reasoning in general and human suspicion in particular. Techniques such as "explaining away" and "data reliability" are powerful in manipulating suspicion and implementing covertness. Additional work is clearly needed for a deeper theoretical understanding of the underlying cognitive process and more comprehensive guidelines for covert intervention delivery in real-world situations.

Acknowledgments

This work was partially supported by an Air Force Office of Scientific Research (AFOSR) grant (FA9550-07-1-0181), and an Office of Naval Research (ONR) grant (N00014-08-1-0042). We would like to thank Scott Thompson and Leanne Hirshfield for the IMP conceptualization.

References

Brachman, R., & Levesque, H. (2004). *Knowledge representation and reasoning*. San Francisco, CA: Morgan Kaufmann.

Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley Publishing Company.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*(3), 340-347.

Fann, K. T. (1970). *Peirce's theory of abduction*. The Hague: Martinus Nijhoff.

Josephson, J. R., & Josephson, S. G. (1994). *Abductive inference: Computation, Philosophy, Technology*. Cambridge, NY: Cambridge University Press.

Paul, G. (1993). Approaches to abductive reasoning: An overview. *Artificial Intelligence Review, 7*, 109-152.

Peirce, C. S. (1931). *Collected papers* (Vol. 1-6). Cambridge, MA: Harvard University Press.

Posner, M. I. (Ed.). (2004). *Cognitive neuroscience of attention*. New York: Guilford Press.

Sun, Y., Wang, H., Zhang, J., & Smith, J. W. (2008). Probabilistic judgment on a coarser scale. *Cognitive Systems Research, 9*(3), 161-172.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, N.J.: Princeton University Press.

Wang, H., & Fan, J. (2007). Human attentional networks: A connectionist model. *Journal of Cognitive Neuroscience, 19*(10), 1678-1689.

Wang, H., Johnson, T. R., & Zhang, J. (1998). UEcho: A model of uncertainty management in human abductive reasoning *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1113-1118). Hillsdale, NJ: Lawrence Erlbaum.

Wang, H., Johnson, T. R., & Zhang, J. (2006). The order effect in human abductive reasoning: An empirical and computational study. *Journal of Experimental and Theoretical Artificial Intelligence, 18*(2), 215-247.

Wang, H., Liu, X., & Fan, J. (2012). Symbolic and connectionist models of attention. In M. I. Posner (Ed.), *Cognitive Neuroscience of Attention* (pp. 47-56). New York: The Guilford Press.