Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems

Madhavan, P. & Wiegmann, D. A.

Human Factors, Vol. 49, No. 5, pp. 773-785

Introduction

- 複雑なシステムで自動化システムの使用が増加
  - 例：飛行機のコクピット，原子炉，トラフィックコントロール
    - 人間と自動化システムの関係性が変化してきている
  - 人間の役割
    - 第一操縦責任者　→　自動化システムのチームメイト
      - 人間は自動化システムと操縦責任を分け合う
  - 特に，近年の自動意思決定システム（Decision support system: DSS）
    - 道具よりもパートナーとして作られている
      (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004)
  - 人間-人間と人間-機械のチームを比較した研究
    - 人々とコンピュータ，ロボット，または，相互作用を行う機械との関係性
      - 人々と他人との関係性に類似している
        (Nass, Fogg, & Moon, 1996; Reeves & Nass, 1996)

- しかし，自動化システムへの信用は，人間への信用とは異なる
  - 人間-人間のチームよりも人間-自動化システムのチームでは
    - パートナーに対する信用が，意思決定プロセスに大きく影響
      (e.g., Dijkstra, 1999)
  - パートナーが人間よりも自動化システムの場合に，初期の信用は高い
    - Perfect automation schema
      - 自動化システムは完璧である
        (Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001)
  - エラーを起こしたパートナーが，人間よりも自動化システムの場合に，信用は大きく低下（Dzindolet et al., 2001）

- 人間，自動化システムへの信用に関する先行研究
  - 人間-人間と人間-自動化システムにおける信用について比較，検討
    - 人間：「前に同様の課題を行った参加者」と教示
    - 自動化システム：「機械(マシーン)」と教示
  - 適切ではない

- ✧ 情報が少なすぎる
- ✧ 実際の状況では，より多くの背景情報が提示される
  - – このような情報が，相手の意見を取り入れるか否かに影響するだろう

- 背景情報の影響について
  - ➢ Lerch, Prietula, & Kulik (1997)
    - ✧ 唯一の実験的検討を行った研究
      - – 人間-人間(エキスパート)，人間-人間(ノービス)のチームを設定
      - – それぞれのチームを，人間-自動化システムのチームと比較
    - ✧ 自動化システムの情報提示は行われていない
      - – 今回の実験では，人間と自動化システムの背景情報を提示

- 信用に関する先行研究
  - ➢ 課題前と後における信用の変化について検討があまり行われていない
  - ➢ 実験の課題前に，参加者が持つ自動化システムへの信用傾向を提示させた場合
    - ✧ 課題前の自らの宣言が，課題のパフォーマンスに影響する可能性
      - – Cognitive anchoring (Madhaven & Wiegmann, 2005)

- 目的 1
  - ➢ 課題前に，人間と DSS パートナーへの信用傾向を，Cognitive anchoring の影響なく実験的に測定し，背景情報による先入観のバイアスが存在するか検討
- 目的 2
  - ➢ 背景情報による先入観のバイアスが，課題における trust calibration[1]と人間または DSS パートナー使用にどのように影響しているか検討

Study1: Assessing a priori biases

- 参加者の人間と DSS パートナーに対する先入観の測定

Method

- 参加者
  - ➢ 大学生 40 名（男性 18 名，女性 22 名，平均年齢 22.5 歳）
  - ➢ 1 時間につき 8 ドルの報酬

- 手順

[1] Trust calibration－自動化システムの性能と自動化システムへの信用との適合性（Lee & See, 2004）

- 参加者には，荷物診断課題を 4 人の異なるアドバイザーと行うと説明
- 参加者に，各アドバイザーの背景情報を提示（Appendix）
  - 人間のノービス
  - 人間のエキスパート
  - 自動化システムのノービス
  - 自動化システムのエキスパート
    - 順序はカウンターバランス
- ２つのアンケートを実施
  - System trust scale (Jian, Bisantz, & Drury, 2000)
    - 1(strongly disagree)～10(strongly agree)
  - 性能評価
    - 10 段階で評定

- 仮説
  - 信用
    - エキスパートと表示された場合
      - 人間がエキスパートである場合，素質，気質に対する信頼性が高くなる (Lerch et al. 1997)
        - 自動化システムよりも人間の方が信用は高くなる
    - ノービスと提示された場合
      - 人間よりも自動化システムの方が，より合理的であると認識される (e.g., Dijkstra, 1999)
        - 人間よりも自動化システムの方が信用は高くなる
  - 性能評価
    - エキスパート，ノービスの両方
      - 性能評価は，パフォーマンスの評価，期待であり，エキスパート/ノービスという気質の影響をあまり受けない
        - 人間よりも自動化システムの方が性能評価は高くなる

Results

- 方法
  - 分散分析の後に事後分析を行う．Cohen's d の効果量を提示.
- System trust scale
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)の分散分析
    - アドバイザーの主効果あり（$F(1, 156)=2.9$, $p=.0009$）

- ✧ 気質の主効果あり（$F(1, 156)=108.87$, $p=.0001$）
- ✧ 2 要因の交互作用あり（$F(1, 156)=10.42$, $p=.008$）
  - – 気質がエキスパートの場合
    - 人間（平均=8.00） ＞ 自動化システム（平均=7.00）
      （$t(78)=1.28$, $p=.01$, $d=0.30$）
      - ▫ 仮説通り
  - – 気質がノービスの場合
    - 人間（平均=5.15） ＜ 自動化システム（平均=6.18）
      （$t(78)=3.06$, $p=.0091$, $d=0.69$）
      - ▫ 仮説通り

- 性能評価
  - ➢ 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)の分散分析
    - ✧ アドバイザーの主効果あり（$F(1, 156)=5.4$, $p=.007$）
    - ✧ 気質の主効果あり（$F(1, 156)=119.41$, $p=.0002$）
    - ✧ 2 要因の交互作用なし（$F(1, 156)=0.12$, $p=.31$）
      - – 気質がエキスパート，ノービスの両方
        - 自動化システム（平均=6.95） ＞ 人間（平均=5.95）
          （$t(78)=2.71$, $p=.043$, $d=0.62$）
          - ▫ 仮説通り

Discussion

- アドバイザーがノービスである場合
  - ➢ 自動化システムの方が信用は高かった
- アドバイザーがエキスパートである場合
  - ➢ 人間の方が信用は高かった
    - ✧ 人間は自動化システムよりも，気質の特徴によって評価される
      (Lerch et al, 1997)

- アドバイザーがエキスパート，ノービスの両方
  - ➢ 人間よりも自動化システムの方が性能評価は高かった
    - ✧ 実際のパフォーマンスを観察した後に，どのように変化するか Study 2 で検討

- 参加者の信用は，背景情報や見かけの信頼性に大きく影響をうける
  - ➢ 疑問

- ✧ 先入観のバイアスは，課題におけるパートナーの診断支持，却下の判断に影響するか？実際のパフォーマンスの方が，判断に影響するか？

Study2: Assessing use of advice and post hoc trust

- 目的
  - Study1 で確認された先入観のバイアスが，実際の行動に反映されるか検討
  - Study1 における課題前の信用と Study2 における課題後の信用を比較，検討

Method

- 参加者
  - 大学生 180 名（男性 72 名，女性 108 名，平均年齢 20.5 歳）
  - 1 時間につき 8 ドルの報酬

- 手順
  - 課題
    - ✧ 空港の荷物検査のシミュレーション
      - – 試行ごとに荷物のレントゲン写真を提示
        - • アドバザーの診断結果を提示
      - – 荷物の中に武器がある/ないを診断
    - ✧ 200 試行実施
    - ✧ 試行ごとに診断結果をフィードバック
      - – hit, miss, false alarm, correct rejection
  - アドバイザー
    - ✧ 人間/自動化システム
    - ✧ 参加者は，アドバイザーの診断を考慮して，診断結果を決定
  - 実験条件
    - ✧ アドバイザー：人間/自動化システム―参加者間
    - ✧ 気質：エキスパート/ノービス―参加者間
    - ✧ 性能：高/低―参加者間
      - – 性能高：90%正確な診断
      - – 性能低：70%正確な診断
        - • 参加者は，実際の性能を知らされない
    - ✧ 20 名はアドバイザーなし
  - 課題終了後に 2 つのアンケートを実施
    - ✧ System trust scale
    - ✧ 性能評価

- 仮説
  - 信用
    - 性能高
      - 人間よりも自動化システムの診断の方が支持される
        - 90%の正解率により，Perfect automation schema が保持
    - 性能低
      - アドバイザーがノービスの場合
        - 人間よりも自動化システムの診断の方が支持される
      - アドバイザーがエキスパートの場合
        - 自動化システムよりも人間の診断の方が支持される
          - Perfect automation schema が崩壊
          - 人間よりも自動化システムのエラーが際立つ

Results

- 方法
  - 分散分析の後に事後分析を行う．Cohen's d の効果量を提示.
- Advice acceptance
  - 全 200 試行を 5 ブロック(40 試行ずつ)に分割
  - 支持の種類
    - Compliance rate：アドバイザーが武器ある，参加者が武器あると診断した割合
    - Reliance rate：アドバイザーが武器ない，参加者が武器ないと診断した割合

- アドバイザーとの診断一致率
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス) ×2(性能：高/低)×2(反応：compliance/reliance)×5(ブロック)の分散分析
    - 性能の主効果あり（$F(1, 152)=35.53$, $p=.004$）
    - 反応の主効果あり（$F(1, 152)=152.73$, $p=.0001$）
    - ブロックの主効果あり（$F(4, 608)=10.42$, $p=.008$）
    - 性能と反応の交互作用あり（$F(1, 152)=5.59$, $p=.006$）
    - 性能とブロックの交互作用あり（$F(4, 608)=6.79$, $p=.0072$）
    - 気質とブロックの交互作用あり（$F(4, 608)=4.09$, $p=.0081$）
    - 気質，性能，ブロックの交互作用あり（$F(4, 608)=3.90$, $p=.009$）
    - アドバイザー，気質，性能，反応，ブロックの交互作用あり（$F(4, 608)=3.18$, $p=.0064$）
  - 性能低よりも性能高で，アドバイザーとの診断一致率が高い

- ✧ 性能高（平均=.51） ＞ 性能低（平均=.41）
  - – Compliance と Reliance に分けて，各性能(高，低)で分析を行う
- ➢ 性能高では，アドバイザー要因や気質要因に関する有意差はみられなかった
  - ✧ 性能低に関してのみ分析を行う

- 性能低の Compliance rate
  - ➢ 2(アドバイザー : 人間/自動化システム)×2(気質 : エキスパート/ノービス)×5(ブロック)の分散分析
    - ✧ 主効果はみられなかった
    - ✧ アドバイザーとブロックの交互作用あり（$F(4, 304)=2.73$, $p=.043$）
    - ✧ 気質とブロックの交互作用あり（$F(4, 304)=3.69$, $p=.003$）
    - ✧ アドバイザー，気質，ブロックの交互作用あり（$F(4, 304)=4.31$, $p=.0072$）
  - ➢ Figure 1 は，性能低における Compliance rate の結果
    - ✧ 気質がエキスパートの場合
      - – ブロック 1～3
        - • 人間と自動化システムの Compliance に差はない
      - – ブロック 4，5
        - • 人間（平均=.58） ＞ 自動化システム（平均=.41）
          （$t(38)=2.70$, $p=.35$, $d=0.88$）
    - ✧ 気質がノービスの場合
      - – 人間（平均=.52） ＜ 自動化システム(平均=.56)
        （$t(38)=1.43$, $p=.067$, $d=0.46$）

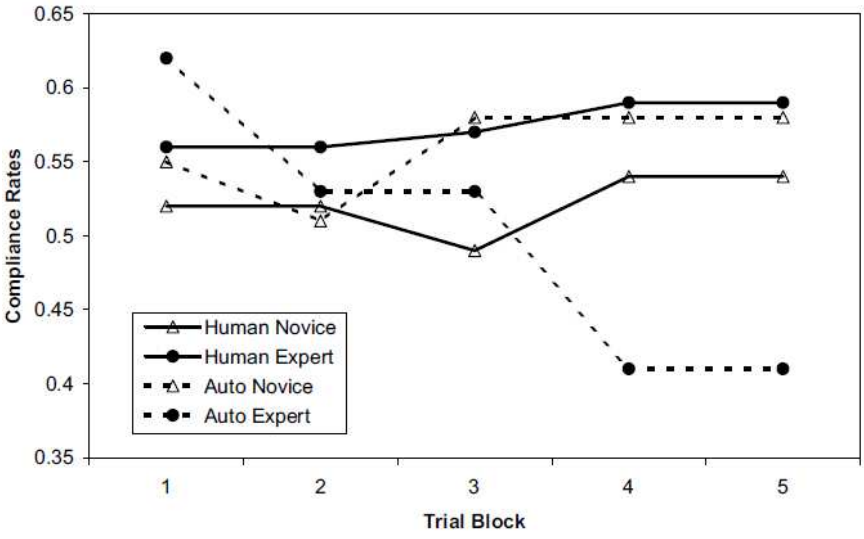

Figure 1. Compliance rates of participants receiving 70% reliable advice.

- 性能低の Reliance rate
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)×5(ブロック)の分散分析
    - ブロックの主効果あり（$F(4, 304)=4.33$, $p=.0033$）
    - 気質とブロックの交互作用あり（$F(4, 304)=1.63$, $p=.025$）
    - アドバイザー，気質，ブロックの交互作用あり（$F(4, 304)=1.88$, $p=.035$）
  - Figure 2 は，性能低における Reliance rate の結果
    - 気質がエキスパートの場合
      - ブロック 5
        - 人間(平均=.84) ＞ 自動化システム(平均=.81)
          （$t(38)=1.66$, $p=.051$, $d=0.54$）
    - 気質がノービスの場合
      - 人間(平均=.78) ＜ 自動化システム(平均=.85)
        （$t(38)=1.86$, $p=.07$, $d=0.60$）

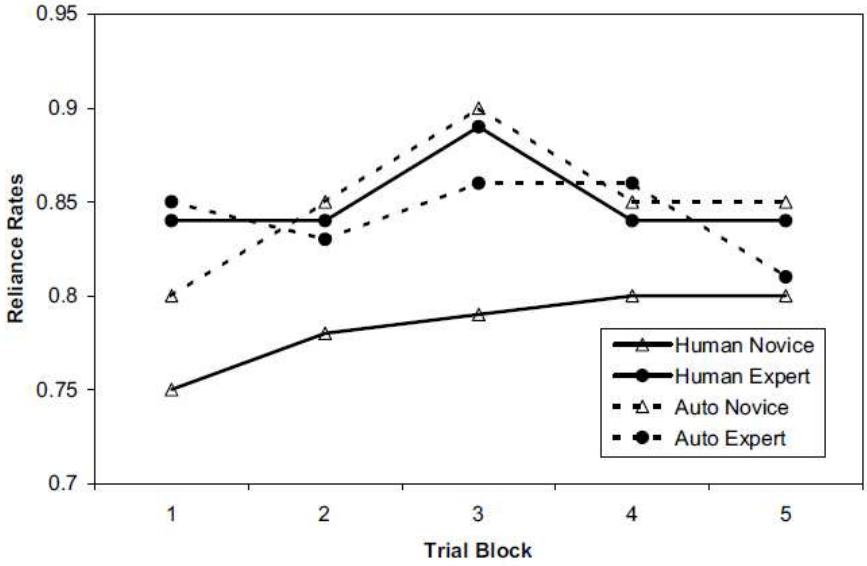

Figure 2. Reliance rates of participants receiving 70% reliable advice.

- Criterion settings (beta)
  - 信号検出理論
    - 参加者が，武器あり/なしのどちらを答える傾向にあったか
    - beta の値が大きいほど武器なしと答える傾向
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)×2(性能：高/低)×5(ブロック)の分散分析
    - 性能の主効果あり（$F(1, 152)=2.28$, $p=.021$）
    - アドバイザー, 気質, 性能, ブロックの交互作用あり（$F(4, 608)=1.19$, $p=.037$）

- 性能高では，アドバイザー要因や気質要因に関する有意差はみられなかった
  - 性能低に関してのみ分析を行う

- 性能低の Criterion settings (beta)
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)×5(ブロック)の分散分析
    - 気質の主効果あり（$F(1, 76)=1.41$, $p=.035$）
    - アドバイザー，気質，ブロックの交互作用あり（$F(4, 304)=1.20$, $p=.042$）
      - エキスパート（平均=1.77） ＞ ノービス（平均=1.48）
        （$t(38)=1.77$, $p=.035$, $d=0.61$）
        - アドバイザーがエキスパートの方が，武器なしと答える傾向
  - 図 3 は，性能低における criterion setting (beta)の結果

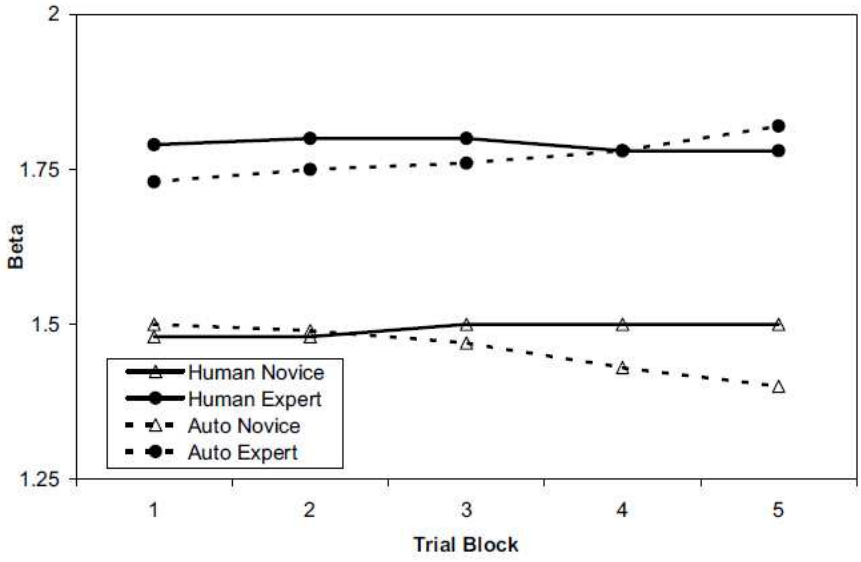

Figure 3. Criterion settings of participants receiving 70% reliable advice.

- Trust ratings
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス) ×2(性能：高/低)の分散分析
    - アドバイザーの主効果あり（$F(1, 152)=3.16$, $p=.0001$）
    - 気質の主効果あり（$F(1, 152)=3.33$, $p=.004$）
    - 性能の主効果あり（$F(1, 152)=43.23$, $p=.0064$）
      - 人間（平均=6.22） ＜ 自動化システム（平均=6.61）
        （$t(158)=1.58$, $p=.038$, $d=0.25$）
      - エキスパート（平均=6.60） ＞ ノービス（平均=6.20）

($t(158)=1.62$, $p=.0063$, $d=0.26$)

- 性能高（平均=7.12） ＞ 性能低（平均=5.71）
  ($t(158)=6.5$, $p=.003$, $d=1.03$)

- 性能評価
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス) ×2(性能：高/低)×2(反応タイプ：hit rate/correct rejection rate)の分散分析
    - 性能の主効果あり（$F(1, 152)=82.76$, $p=.002$)
      - 性能高（平均=82.5%） ＞ 性能低（平均=62.55%）
        ($t(158)=9.18$, $p=.0042$, $d=1.46$)

- 性能高の性能評価
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)の分散分析
    - 気質の主効果あり（$F(1, 76)=1.32$, $p=.046$)
      - エキスパート（平均=84%） ＞ ノービス（平均=78.3%）
        ($t(38)=1.39$, $p=.033$, $d=0.45$)

- 性能低の性能評価
  - 2(アドバイザー：人間/自動化システム)×2(気質：エキスパート/ノービス)の分散分析
    - アドバイザーと気質の交互作用あり（$F(1, 76)=1.08$, $p=.033$)
      - アドバイザーが人間の場合
        - エキスパート（平均=67.3%） ＞ ノービス（平均=59.9%）
          ($t(38)=1.54$, $p=.025$, $d=0.50$)
      - アドバイザーが自動化システムの場合
        - エキスパート（平均=60.07%） ≒ ノービス（平均=67.6%）
          ($t(38)=0.82$, $p=.12$, $d=0.27$)
          - 優意差はないが，ノービスの方が性能評価は高い

Discussion

- Behavioral dependence on advisers
  - 性能が高い場合
    - 背景情報に関わらず，参加者はアドバイスに従った
      - 完璧に近いパフォーマンス(90%の正解率)の影響
  - 性能が低い場合

- ✧ 人間のエキスパートの診断は支持された
- ✧ 自動化システムのエキスパートの診断は，最初は支持され，後半支持されなくなった
  - – perfect automation schema の崩壊
    - • 仮説通り

- Decision criterion setting
  - ➢ 性能が低い場合
    - ✧ 背景情報によって，beta の値が異なる
      - – エキスパートよりもノービスで，beta の値は大きかった
    - ✧ ノービス
      - – Study 1 と同様にアドバイザーへの期待が低かった
        - • アドバイスの影響を受けず，hit を増やし，アドバイザーの miss を削減しようと考えた
    - ✧ エキスパート
      - – Study 1 と同様にアドバイザーへの期待が高かった
        - • アドバイスに従い，false alarm を削減しようと考えた
    - ✧ 背景情報によって，アドバイザーの診断支持方略が異なる

- Subjective trust and perceived reliability
  - ➢ 信用に関して Study 1 との比較
    - ✧ 性能低
      - – Study 2 の結果は Study 1 の結果とは異なる
      - – 先入観のバイアスと実際のパフォーマンスを観察した後に持つバイアスは異なる
        - • 信用は，実際のパフォーマンスから強い影響を受ける

General discussion

- Madhavan & Wiegmann (2007)
  - ➢ DSS に人間らしい特徴を持たせる
    - ✧ DSS は人間と同様に信用されるのではないか
- 今回の実験
  - ➢ 人間と自動化システムという事前情報を提示
  - ➢ 同一の課題パフォーマンスを示す状況
    - ✧ 課題を行う前の時点で，DSS と人間への信用は異なった
      - – 人間らしい特徴を備えても，DSS は，人間と同じようには信用されない

のでは？

- DSS を人間に近づけるよりも，DSS と人間との違いをオペレータに理解させた方が，有効な DSS 使用につながる可能性

Conclusion

- DSS に備えられた気質
  - 背景情報として提示されることによって，DSS への信用，依存に影響する
- DSS をデザインする際
  - DSS に関する情報提示には注意が必要

APPENDIX

The Novice Human Adviser

"Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from a novice student named BILL JOHNSON. BILL JOHNSON is currently an undergraduate student and has no prior experience in real world luggage screening tasks. He is currently majoring in criminology at a small technical college in the Midwest and is interested in specializing in antiterrorism and airport security. However, BILL is still a novice when it comes to modern terrorist tactics. He possesses limited knowledge of the types of modern weapons and explosives commonly smuggled aboard aircraft. He also has a partial understanding of the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. BILL has recently applied for an internship at the Transportation Safety Administration (TSA) to help oversee security operations."

The Novice Automated Adviser

"Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from a novice computer system called DETECTOR, which is an automated diagnostic aid that has been designed to identify hidden contraband in airline passenger luggage. DETECTOR is based upon the technology traditionally used at major airport security checkpoints over the past 10 years. DETECTOR was designed and developed at a small technical college in the Midwest, which contains a recently established department in antiterrorism and airport security. It currently possesses a limited database of the types of modern weapons and explosives commonly smuggled aboard aircraft. Its algorithms are relatively unsophisticated in their attempts to capture the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. The Transportation Security Administration (TSA) is considering whether to conduct a limited field test of DETECTOR at a small airport in the hope of making it employable at larger airports to enhance security operations in the future."

The Expert Human Adviser

"Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from an expert in airport security named DR. BILL JOHNSON. DR. BILL JOHNSON was originally trained as a luggage screener, serving 10 years in some of the busiest airports in the United States. He went on to earn his Ph.D. in criminology from the Massachusetts Institute of Technology (MIT), specializing in antiterrorism and airport security. DR. BILL JOHNSON is an expert when it comes to modern terrorists' tactics. He possesses extensive knowledge of the types of modern weapons and explosives commonly smuggled aboard aircraft. He also has a keen understanding of the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. DR. BILL JOHNSON has recently been appointed by the Transportation Security Administration (TSA) to oversee security operations at Chicago's O'Hare International Airport, which is one of the largest airports in the world."

The Expert Automated Adviser

"Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from an expert computer system called SUPER-DETECTOR, which is an automated diagnostic aid that has been programmed to identify hidden contraband in airline passenger luggage. SUPER-DETECTOR is based upon, yet far exceeds, the technology traditionally used at major airport security checkpoints over the past 10 years. SUPER-DETECTOR was designed and developed at the Massachusetts Institute of Technology (MIT), which contains a highly specialized department in antiterrorism and airport security. It possesses an extensive database of the types of modern weapons and explosives commonly smuggled aboard aircraft. Its algorithms are highly sophisticated and effectively capture the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. SUPER-DETECTOR has recently been employed by the Transportation Security Administration (TSA) to enhance security operations at Chicago's O'Hare International Airport, which is one of the largest airports in the world."