日産科学振興財団研究報告書，24(2001)

信頼の認知モデルに基づくヒューマンマシンシステムのリスク制御

Modeling of Trust in Machines for Risk Management of Human-Machine Systems

研究代表者 伊藤 誠
電気通信大学 大学院情報システム学研究科 助手
Makoto ITOH, Research Associate
Graduate School of Information Systems, University of Electro-Communications

This study proposes a new model of trust. In this model, trust is the combination of the range of expectation of successful work of an automation and the level of willingness to rely on the automation (LWRA). Mistrust, such as false distrust or false trust, are special cases of this model. We show that the Dempster-Shafer theory can be applied to describe updating of trust in terms of the LWRA, and clarify that the ambiguity about trustworthiness of automation plays a role of a buffer. We also show that lack of information on how an automation works and on the limit of the automation is one of causes of being false trust.

1．研究目的

自動化システムへの信頼感(trust)の誤り，即ち過信・不信による事故やトラブルの防止策確立を目指す．本研究では，特に過信に注目し，trustの認知モデルを構築し，trustの変化及び過信への過程を明らかにする．

2．研究経過

2.1 Trustとシステム安全性の関係の明確化

研究を始めるにあたり本学の鈴木和幸教授のレビューを受けることができ，次のコメントを得た．すなわち，品質管理の分野では，最も多く不具合を起こしている問題から対策を立てるのが常であるが，この観点からすると，trustと安全性との関係を吟味すべきである．実際，両者の定量的な関係は十分な議論がない．

そこで，事例収集が容易な航空分野に注目し，事例データベース(DB)を調査した．NASAのASRSによれば，約20万件の事例中，ヒューマンファクターに起因する事例が70%以上を占めている．しかし，ASRSのデータはそれ以上を詳細に分類できる形になっていない．そこで，Flight Deck Automation (http://flightdeck.ie.orst.edu/fdai/issues.html)を分析した．その結果，自動化に関連した事故や事例のうち，約10%が過信の問題を含んでいた．この値は一見小さいが，100以上の全分類項目中，過信を意味する2項目が約10%を占めていることは，自動化への過信が事故を引き起こしやすいことを示唆していると考えられよう．

さらに，安全確保のための自動化システムを過信しやすい状況を作り出すと，自動化により生じる安全余裕を，人間が生産性向上を得ようと悪用することを，認知実験により確かめた[3]．この悪用により，自動化が導入されても危険性が減少しない傾向があった．

以上から，過信が安全性に重大な影響を及ぼすことを確認できた．

2.2 従来のtrust研究の知見の整理[5]

つぎに，日本ではtrustを体系的に論じた論文がないため，従来のtrust研究を調査し，知見を整理した．

たとえば，trustには時間的な側面がある．すなわち，相手の行動の予測可能性（Predictability:P)，Pが得られた上での相手への依存可能性（Dependability:D)，Dが醸成した上での未知の状況での相手への依存可能性（Faith:F)の3つである．

また，通常，P,D,Fおよびトータルとしてのtrustについて，10段階のスケールで主観評価させ，その変動を数理モデルで記述する．しかし，この方法では，信頼感の変動は記述できるが，このtrustと過信・不信との関係が不明確であることが明らかになってきた．

2.3 期待の範囲としてのTrustモデルの構築

2.3.1 trustの構造モデル[4]

過信・不信との関係を明確にするtrustモデルを構築した（図1）．このモデルでは，自動化システムの実行

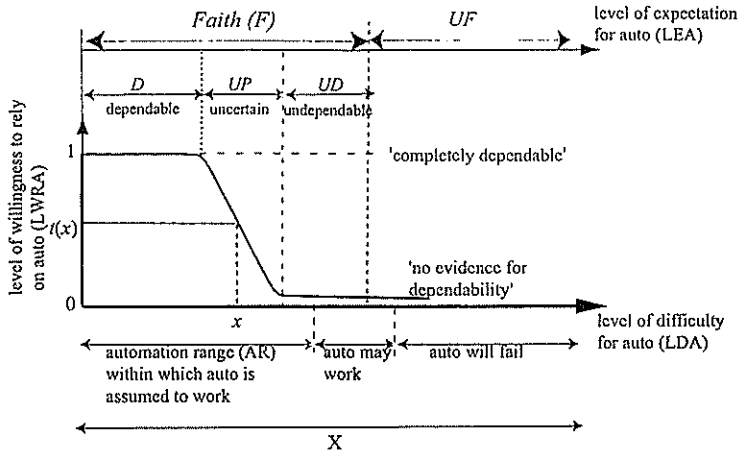


図1　Trustの構造

環境により定まる実行の困難さ(LDA)を導入している点に特徴がある．このことにより，trustの時間的側面(D,P,F)を明示的に反映できている．LDAを導入する理由は，過信による事故の多くで，本来自動化システムが動作しない状況において自動化へ依存しようとしたことが原因となっているものが見られるからである．

図1のモデルでは，trustを，期待の範囲（横軸)と依存性のレベル（縦軸)の組合せと考える．LDAに関し，自動化システムのタスク成功を期待する範囲(F)のうち，信頼できる部分(D)と信頼できない部分（UD)の間に，信頼できるか否か定かでない部分（UP）がある．また，縦軸の自動化システムへの依存性(LWRA)に関し，Dの範囲内では1，UDの範囲内では0とする．UPの範囲内では0から1の間の値を取る．

2.3.2 trustの変化モデル[2]

Trustの変化は，図1においては，縦方向の変化と横方向の変化がある．

縦方向のtrustは，従来研究におけるtrustの主観評価に相当する．このtrustは，一次遅れ系で記述できることが知られている．また，自動化システムの誤動作によりtrustが変化する場合，誤動作発生回数がトータルで一定であっても，ある一時期に集中的に起こるのと，時間をおいて散発的に起きるのとでは，trustの低下の程度やその後の回復の仕方が異なること（誤動作発生パターンへのtrustの推移の依存性）が，筆者が以前に行った実験の結果知られていた．しかし，その理由は従来の一次遅れ系モデルでは説明できない．

本研究では，誤動作発生パターンへのtrustの変化の依存性を説明する為に，trustの縦方向の変化を，Dempster-Shaferの証拠理論によってモデル化した．

ある実行状況xにおいて，自動化システムは{信頼できる（Tx), 信頼できない(UTx)}のいずれかである．Ωx={Tx,UTx}を「信頼できるか判断できない」ことを意味する命題とする．関数mを，{Tx},{UTx}, Ωxそれぞれの命題に対する確信度の和が1になるように与えるものとする．

いま，オペレータのtrustが，図2の左側にあるように，m1({Tx})=a,m1({UTx})=b,m1(Ωx) =1-a-bであり，自動化の誤動作を経験したものとする．この誤動作は，m'({UTx})=c,m'(Ωx)=1-cとして表現されるとする．このとき，事前の信念m1は，Yagerの規則による情報統合に基づいて，m'により更新される．この際，{Tx}に割り当てられていた確信度の一部が，Ωxに移動する（図2)．このことは，「不明」を意味するΩxがバッファとして機能する，すなわち，誤動作を経験しても「信頼」が直ちに「不信頼」に結びつくのではないことを意味する．

たとえば，m1({Tx})=1，すなわち，オペレータが自動化システムを完全に信頼している状況を考える．ここで，自動化の誤作動により，m'({UTx})=0.7,m'(Ωx)=0.3を得たものとすると，オペレータの信念m1は，m1(Tx)=0.3,m1(Ωx)=0.7と更新される．

この，「不明」部分がバッファとして機能することが，誤動作発生パターンへのtrustの依存性をもたらす理由となっている．図3は，50回の人間と自動化システムと

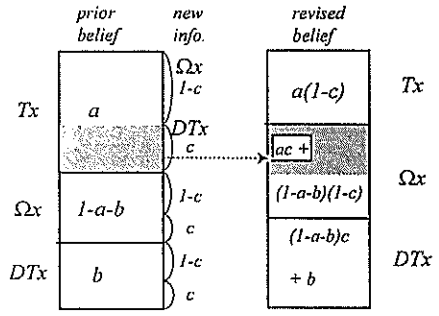


図2　誤動作を経験することによるtrustの低下

の協働において，3回の誤動作が起こる場合の，trustの推移をDS理論に基づくモデルで計算したものである．誤動作が一時期に一度だけ発生すると，一時的に信頼感が「不明」感に変わるが直ちに元に戻る．しかし，連続的に誤動作を経験すると「不明」感がさらに「不信」感へと移行する（図3）．図3におけるtrustの推移は，実験で観察されたデータとよく似ている．

2.4 認知実験：過信への傾向の検証[1]

2.4.1 実験の目的および仮説

図1の縦軸方向，すなわち依存性のレベルについては，証拠理論的モデルにより，どのように変化するかを説明できた．trustの変化の原理を理解するためには，つぎに，期待の範囲の変化がいかなる要因によっておきるかを明らかにする必要がある．その要因には様々なものが考えられるが，本研究では自動化システムの動作限界の情報に注目する．

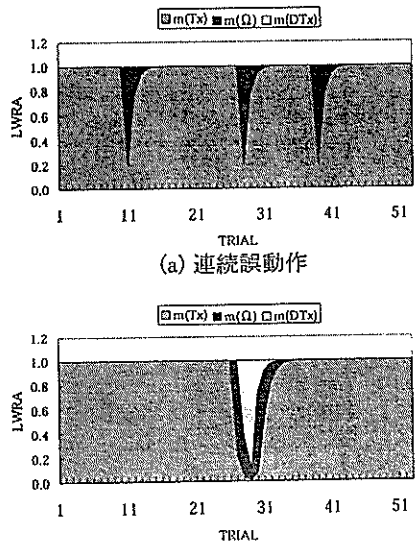
図1にあるように，自動化システムの動作の限界には2種類ある．一つは設計上の限界（安全限界）であり，仕様として明確に与えられる．安全限界を超えた状況でも自動化システムがタスクを成功させることはある．いま一つの限界は，真の限界（危険限界）であり，これを超えるとほぼ確実に自動化はタスクを成功できなくなる．ここで，3つの仮説を立てる．

（1）安全限界のみを知らされていると，真の限界がわからないために，経験をつむにつれて期待の範囲が広まっていき，場合によっては危険限界を超える

（2）安全限界に加え，危険限界も知らされていると，真の限界を超えなければ問題ないとの判断がなされ，自動化への依存が安全限界をこえ，危険限界まで近づく．また，経験を積むに従い，危険限界を安全限界と錯覚し，期待が危険限界を超えうる

（3）危険限界に加えその根拠が知らされている場合，期待の範囲は，危険限界まで広がる反面，危険限界を超えることはない

本研究では，以上の仮説を認知実験で検証する．



(a) 連続誤動作

(b) 離散的誤動作

図3　3回の誤動作経験によるtrustの推移

2.4.2 実験の方法

本実験では，ミックスジュース製造プロセスの仮想プラント制御シミュレータを用いる．一連の作業は自動化されているが，加熱殺菌の設定部には手動操作の余地がある．各バッチにおいて，製造予定量が決められており，それに応じて原料タンクから混合タンクへ原料が送られる．実際の流入量は，設定量より多いことがあるが，誤差が原料比5%以内（危険限界）であれば，自動化による設定で成功しうる．しかし，3%までは，自動制御がほぼ完全に成功するため，供給誤差3%を自動化の設計上の許容範囲とする（安全限界）．オペレータは流入量を監視し，許容範囲を供給誤差が上回る場合には手動介入しなければならない．

本実験には，学生12名が参加している．被験者をグループA,B,Cに無作為に分け，グループ間で与えられる情報が異なるようにする．Aには，安全限界の値（3%）のみを教示する．Bには，安全限界値に加え，危険限界値（5%）を教示する．Cには，安全限界値，危険限界値に加え，その根拠を教示する．

実験は，各被験者に対して，1回約1分の試行を一日につき100試行，これを3日間実施した．データ採集に先立ち，実験の目的，内容を教示した．つづいて，練習として，タスクがいかに進むか，どのようにして介入すればよいかを試行しながら確認させた．

本実験では，自動化システムへの信頼感を，モードの境界，委任率，思考時間を測定する事により総合的に推定する．なお，ここでは紙面の都合上，モードの境界に限定して結果と考察を示す．

オペレータは，供給誤差の絶対値が十分に小さければ自動化に依存するが，供給誤差が大きくなると自動化にとってタスクが困難となるため，ある値の供給誤差以上ではオペレータが手動介入するようになる（図4）．この，自動化への依存と手動介入との境にある供給誤差の値を，モード境界値とよぶ．図4で明らかなように，モード境界値は精確な値として得られる．

2.4.3 結果と考察

図5に，各被験者のモード境界値の推移を示す．

すべての試行において自動化へ依存する被験者がいたため，各グループにおけるモード境界値の平均値を議論することには意味がない．しかし，モード境界値の精確さに基づけば，全体的には次のことが言える．

・グループAでは，経験が増すにつれてモード境界値

が上昇する傾向がある

・グループB，Cでは，経験が増すにつれてモード境界が上昇するとは断定できない

・グループBでは，モード境界値が危険限界側に偏る傾向がある

・グループCでは，モード境界値が安全境界側に偏る傾向がある

こうした考察の結果として，2.4.1節に示した仮説は支持されたといってよい．自動化システムの動作原理への理解不足は，過信につながりやすいといえる．

3．研究成果

本研究の結果，次の2点が明らかになった．

- trustの変化において，依存できるか否か不明である部分がバッファとして機能する．このことは，誤動作が起きても信頼感を損なわずにすむ反面，システムの機能低下の兆候を見逃す可能性があるという意味で，過信につながるメカニズムとしても働きうる
- 自動化システムの動作限界や動作原理への理解不足は過信をもたらしうる．過信を防ぐには，自動化の安全限界，危険限界のみならず，その根拠をオペレータに与えることが有効である

4．今後の課題と発展

自動化システムの動作原理とその限界は，十分には明らかではないか，もしくはオペレータにとって十分に理解できないことがある．そこで，過信を防ぐための別の方法を解明する必要もある．そのためには，過信をもたらす他の要因とその影響を明らかにしなければならない．

また，図1における横軸方向のtrustの変化，とくに過信への傾向について，定量的に予測できる数理モデルを構築する必要がある．この数理モデルにより，過信を防ぐためにはどの要因から対策を立てればよいかを明確にできると期待される．

さらに，本研究では，自動化システムに対するオペレータのtrustと過信・不信を考察しているが，同様の議論は，組織における過信の問題に拡張できる．この数年の組織の不祥事をみると，システムや制度への過信が多く見られる．本研究で得られる知見は，こうした組織の過信を防ぐための指針を打ち立てられる可能性を秘めている．

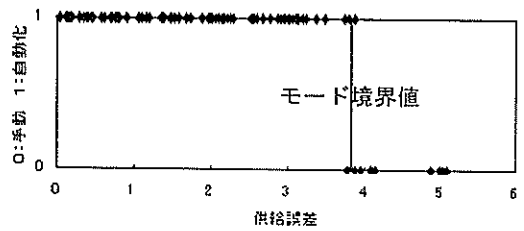
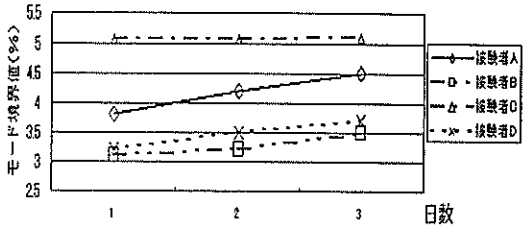
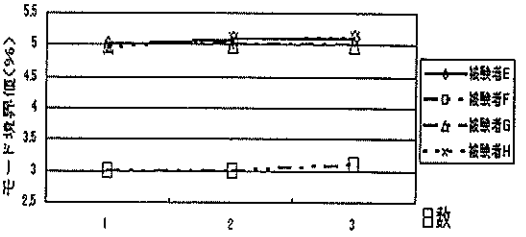


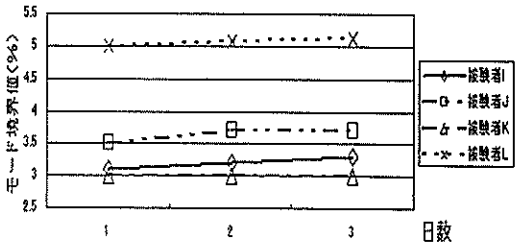
図4　グループ1被験者Aのモード境界（第1日目）



(a) グループA



(b) グループB



(c) グループC

図5　モード境界値の推移

5．発表論文リスト

[1] 伊藤（投稿準備中).「自動化システムにおける情報の与え方の違いによるオペレータの過信の分析」

[2] Itoh, M. (2001). "Ambiguity as a Buffer in Change of Trust in Automation", *Proc. 8th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, Kassel, Germany (to appear).

[3] Itoh, M., Sakami, D., and Tanaka, K. (2000). "Dependence of Human Adaptation and Risk Compensation on Modification in Level of Automation for Systems Safety", *Proc. IEEE-SMC2000*, pp. 1295-1300.

[4] Itoh, M., and Tanaka, K.(2000). "Mathematical Modeling of Trust in Automation: Trust, Distrust, and Mistrust", *Proc. IEA2000/HFES2000*, pp. 1-9 - 1-12.

[5] 伊藤，田中(2000).「ヒューマンマシン間の信頼と過信・不信の認知過程」，計測自動制御学会システム情報部門シンポジウム，pp. 37-42.