信頼と権限：　ヒトと機械の共創の視点から

稲垣敏之
筑波大学大学院システム情報工学研究科

クルマの運転においては,「状況の中でなすべきタスク」がつぎつぎに現れる. それらのタスクのなかには, ヒトが単独を実行するには困難が伴うものがある．そこで，機械にも能力の限界があるとはいえ，機械がヒトを補佐する形態が不可欠となる．すなわち，クルマの運転ではヒトに主権が与えられているが，ヒトの認知・判断・操作の完全性は保証できない中でヒューマンファクターに起因する事故を防止するために，機械がヒトを補佐することによって「１＋１＝１」を確実に実現しなければならないという側面がある．一方，ヒトとクルマの協調によって，「１＋１＝３」のような相乗効果が期待できるケースもある．

ヒトと機械が協調するシステム形態において慎重な検討が必要な課題の例として，信頼（過信，不信）に関わる問題と，ヒトと機械の権限の共有・委譲の問題がある．本講演では,「運転支援システムを導入すると過信が生じるのではないか」との議論に潜む落とし穴と,「ヒトに優しいシステム」にするとの名目のもとで機械がヒトの意向に逆らうことは許容されるかといったテーマを取り上げ，ヒトと機械の共創を考察する．

Issues of Trust and Authority for Synergy of Humans and Machines

Toshiyuki INAGAKI
School of Systems and Information Engineering, University of Tsukuba

Human-centered automation is an approach to realize a work environment in which humans and machines collaborate cooperatively. It is usually claimed that, “the human must have final authority over the automation.” However, we ask ourselves whether the statement must hold at all times and on every occasion. This paper argues that authority trading from humans to automation may be indispensable for assuring safety of human-machine systems, and that a machine-initiated automation invocation may be required even in the human-centered automation. This paper also discusses humans may not always trust in machines overly issues even when the machines are smart and reliable.

Keyword: Authority and responsibility, function allocation, human-centered automation, human-machine interaction, risk and safety, situation awareness, trust and over-trust

---

1. 状況認識

交通移動体の運転や操縦は，認知・判断・操作の繰り返しである．自動車の運転を例にとると，自分の車の周辺に他の車や歩行者がいないか，路面が濡れたり凍結したりしていないか，道路形状が急に変わるところはないかなどに気を配り，他車や歩行者に気がついたときはそれらの動静を予測しながら，安全運転の継続のために何をしなければならないかを考え，状況に最も適していると思われる操作を実行する．これが，認知・判断・操作のサイクルの一例であるが，このサイクルの中で最も基本となるのが認知である．すなわち，認知が正しくなければ，それに引き続く判断や操作は正しくありようがない．

自分の置かれた状況の把握という意味での認知と

判断（の一部）を表すことばに「状況認識」がある．状況認識（situation awareness）については，表１に示す３つのレベルを識別するのがふつうである[1]．

表１　状況認識の３つのレベル

レベル１：何かが起こっていることに気づく
レベル２：その原因を同定できる
レベル３：これからの事態の推移が予測できる

まず，何か変だということに気づく（レベル１）．そして，その現象をもたらしている原因が何であるかを特定する（レベル２）．さらに，今，ある行動を取ればどのような結果がもたらされるか，その行動を取らなければどのような事態に移っていくかを予測する（レベル３）．そこまで達成できれば，状況認識は完璧なものとなり，眼前の状況に対して，有効かつ適切な行動を選択し，それを適切なタイミングで実行することができる．すなわち，レベル３までの状況認識が達成できていることが，合理的な意思決定の前提となる．

2. 状況認識の失敗

2-1　レベル１の状況認識の失敗

「何か変なことが起こったら，それに気づく（レベル１の状況認識）くらいは簡単だろう」と考えたいところであるが，「気づき」は意外に難しい．現在の社会には，ヒトと知能機械が共存するシステムが少なくないが，そのようなシステムにおけるレベル１の状況認識の失敗（気づきの失敗）には，知能機械が持つ高い自律性，ヒューマン・インタフェース設計の不備，知能機械の能力の過大評価，環境変化に対する警戒心の欠如などが関与している[2),3)]．

2-2　レベル２の状況認識の失敗

「異常が発生していることには気づいたものの，何が異常の原因であるかが分からない」というケースがレベル２の状況認識の失敗（原因特定の失敗）である．原因特定の失敗には，つぎのようないくつかのタイプがある[2),3)]．

（１）ある現象が起こったとき，それを引き起こす原因についての知識が十分あるにもかかわらず，「現象→原因」の形で記述される診断ルールのうち，その場面に該当しないものを適用する．

（２）眼の前で起こっている「現象」が，少なくともその人にとっては今までに経験したことがない「未知の現象」であり，何を意味しているのかよく分からない．知能機械の「ものの見方・考え方」がヒトと異なることによって発生するオートメーション・サプライズも，このタイプである．

（３）眼前の（奇妙な）現象に対して，ある意味で「合理的」と思える（実は誤った）説明をつけることによって自らを納得させる．これは，ヒトの特性に密接に関わるものである．知能機械からのメッセージを都合良く解釈したり，メッセージを無視したりするケースが該当する．

2-3　レベル３の状況認識の失敗

眼前で起こっている異常に気づいており，その原因がどこにあるかもわかっているにもかかわらず，そこに潜むリスクを過小評価し，「今はまだ対応しなくても大丈夫だろう」と思ってしまうケースは，レベル３の状況認識の失敗（予測の失敗）である．知能機械とのインタラクションのなかで，これから自分が取ろうとする行動がどのような事態をもたらすのかを的確に予測できないときにも起こり得る[2),3)]．

3. ヒトへの支援の時間的多層構造化

交通移動体の安全確保には適正な状況認識が不可欠であるが，ヒトが状況認識を適正に保ち続けることは容易ではない．このことを念頭に置くと，運転・操縦にあたるヒトへの支援機能に表２のような時間的多層構造を持たせることの重要性がわかる[4)]．

表２　支援の時間的多層構造

第１層：	ヒトの状況認識力の強化や「情報提供」．
第２層：	運転に対して障害あるいは危険をもたらす可能性があるものが検出されたとき「注意喚起」．
第３層：	直面する状況の中で必要と考えられる操作が行われていないことを何らかの技術で検知したとき，当該操作を求める「警報提示」．
第４層：	警報提示にもかかわらず当該操作が行われない（遅れている）ことが検知されたとき，機械が当該操作を代行する「制御介入」．

（注）「注意喚起」は，ハザードの存在をヒトに知らせるものであり，ヒトに特定の操作・行動を指示するものではない．「警報」は，ただちに特定の操作あるいは行動をとることをヒトに求めるものである．

表２に示した情報提供，注意喚起，警報提示，制御介入の４つは，国土交通省先進安全自動車（ASV）

プロジェクトで議論されている「４つの支援レベル」に対応している．ただし，ASV プロジェクトでは，個々の支援技術を論じられているが,「時間的な流れの中でヒト（ドライバ）はどのように支援されるべきか」は，ほとんど論議されていない．

4. 「制御介入」の作り込みへのためらい

制御介入については，一部に実用化された技術もあるものの，一般には，そのレベルまで踏み込んだ支援を提供することには躊躇が見られる．その背景には，つぎのような問題がある．

（１）ソフトウェア／ハードウェアの信頼性
（２）運転に関する「ドライバ主権」の考え方
（３）ドライバの心に生ずる「過信」への恐れ

（１）の「信頼性（reliability)」の問題は自明であるので，ここでは取り上げない．

（２）の「ドライバ主権」は,「運転はあくまでもドライバが主体となって行うものである」との主張である．これは，いわゆる「人間中心の自動化」に沿った考え方である．例えば，航空分野における人間中心の自動化では,「運航安全に関する責任は人間にある．したがって，責任を負っている人間には，最終決定権を与えておかなければならない」と考えられている．しかし，航空機以外のシステム（交通移動体の例でいえば，自動車，鉄道，船舶など）に対しても，航空機の場合と同じ考え方で臨んでよいとは限らない．すなわち,「人間中心の自動化」のありかたは，対象となるシステムの特性や，それを操作する人間に対してどれほどの教育・訓練を仮定できるかなどの様々な要因に依存することに注意しておく必要がある．例えば，自動車の場合,「いついかなる場合でもドライバに決定権を与えておく（あるいは押し付けておく)」デザインは，実は，ドライバや周囲の環境（他車，歩行者など）の安全確保に有益ではないことがある．

（３）の「過信」については，次節で考察する．

5. 過信

5-1 信頼の４つの次元

Lee & Moray [5] は，ヒトがシステム（機械）に対して抱く信頼（trust）は，ヒトに対する信頼と本質的に同じであるとし，表３に示すような４つの要件を示した．すなわち，４つの要件のうちに満足されないものがあれば，システムへのヒトの信頼は磐石とはいえない．

表３　信頼の４つの次元

（１）基礎： 自然界を支配する法則や社会の秩序に合致している
（２）能力： 終始一貫して安定的かつ望ましい行動や性能が期待できる
（３）方法： 行動を実現するための方法，アルゴリズム，ルールが理解できる
（４）目的：上記の背後にある意図，動機が納得できる

物理法則に従っている「工学的システム」では（１）は成立していると考えてよい．（２）－（４）をわかりやすく表現するなら,「つねに一貫した動作を反復するものであっても，それを支える論理が誤っているものは信頼できず，また，たとえ論理的な誤りはなくても，正しい目的意識に支えられていると思えないものは信頼できない」ということができる．

たとえば,航空機の空中衝突を防止する TCAS は,（４）は満たすが,（２）は万全ではない．不要あるいは不適切と考えられる回避アドバイザリが発せられることも皆無ではないからである．

オートパイロットやオートスラストなどを駆使して航空機をみごとに操るコンピュータも，パイロットたちに,「いったいこいつは何をしているのだ？」,「つぎは何をやるつもりだ？」と言わせることがある．コンピュータの「意図」が人間に理解できないためである．このようなケースでは,（４）は満たされていない．

表２の要件が客観的には満足されているにも関わらず,「満足されていない」と誤った判断をする場合を「不信」(distrust）と称する．一方，客観的には満足されていないはずの要件に対して「満足されているはずだ」との判断を下す場合を「過信」（over-trust）という．いずれも,「不適切な信頼」（mistrust）の例である [6].

なお,「依存」(reliance）という用語は，ときに「システムに依存するようなことがあってはならない」というように，過信と同じ意味合いで用いられることがあるが，この用法は正しくない．「信頼に値するものに対する依存」は当然の行為あるいは判断であり，ワークロード軽減の観点からも正当なものである．あってはならないものは,「信頼に値しないものに対する依存」である．「信頼」に対して適否を区別したように,「依存」に対しても適否を区別することが肝要である．

5-2　「過信」の多様性：　何に対する過信？

「過信」は,「対象に対する過大な信頼」であるといえるが,「何を対象としているか」によって，過信の形態は異なる．過信のパターンの例を，C-SHEL モデル[6]（図１）に沿って以下に示す.

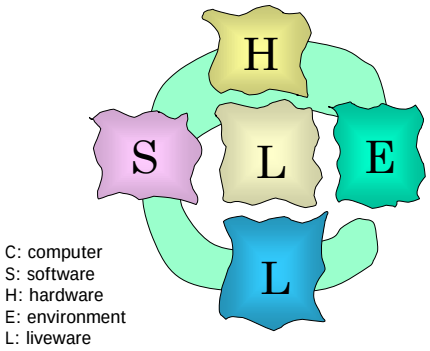

図1　C-SHEL モデル

（１）C への過信の例：ACC を作動させての走行中に,「隣接車線に割り込みの素振りを見せる車がいるが，私の車の ACC システムは，すでにその可能性も考えて制御してくれているはずだ」と考えるケース.
（２）S への過信の例：「マニュアルを読まずに試行錯誤や思いつきで操作しても，システムが壊れたり，危険な（あるいは食い止められないような）振る舞いをしたりすることはないだろう」と考えるケース.
（３）H への過信の例：「この機械は丈夫で壊れないという実績がある．本来なら定期点検を行輪なければならない時期だが，繁忙期でもあるので今回の点検は見送ろう」と考えるケース.
（４）E への過信の例：環境は時間的に変化するものであり，いつ何が起きるかわからないものだが，「ここは，もともと交通量が少ないうえに道路形状も単調で，ほんとうにリラックスできる気持ちのいい場所だ」などと考えながら，まったく警戒心を抱かずにのんびりドライブしているケース.
（５）L への過信の例：前方左手から交差する道路の一時停止線のところに赤い車が停止しているのが見えたとき,「私の車は，赤い車のドライバからも見えているはずだ．何よりも，私が走っているのは優先道路なのだから，赤い車は私の車をやり過ごしてから発進するはずだ」と考え，それ以降，もはや赤い車に注意を向けようとしないケース.

過信の対象が多様であることは,「過信を防止するにはどうすればよい？」との問いに対する答も一様でないことが分かる．例えば,「C への過信」を抑止しようとすれば，つぎの（１）～（４）のように，ヒューマン・インタフェース設計への対策が主要になるが，S, H, E, L への過信を抑止する手段は，また別のものとなる.
（１）機械がなぜそのように判断したのか，その根拠が分かる情報提示.
（２）機械が今何をしようとしているのか，意図理解の手がかりとなる情報提示.
（３）ヒトと機械が状況認識を共有できる情報提示.
（４）機械への過信を防ぐために，機械の能力限界を知る手がかりとなる情報提示.

5-3　運転支援システムの導入は過信を招くか

運転支援システムといっても，平時に作動するものもあれば，緊急時に作動するものもある.「平時に作動するシステム」は，自動車でいえば ACC，航空機でいえばオートパイロットなどに典型を見ることができる．一方,「緊急時に作動するシステム」の例としては，自動車では被害軽減ブレーキ（PCS：Pre-Crash Safety）システム，航空機では TAC（Thrust Asymmetric Compensation：離陸滑走中のエンジン故障によって生じるヨー・モーメントを自動的に補正する機構）[2] が典型である.

平時に作動するシステムは，さまざまな機能を提供することによって，ヒトの負担軽減を図ろうとするものが多い．ヒトがこれらのシステムに「達成すべき目標（ゴール）」を指示すれば，システムは状況をセンシング・理解したうえで，ゴールを達成するのに必要な操作を自律的に実行する能力を持っている．これらのシステムが知的なものであればあるほど，ヒトはそのシステムを信頼し，依存するようになるのは自然なことである.

すでに述べたように,「ヒトがシステムに依存するのは良くない」と考えるのは必ずしも正しくない.「適正な信頼に基づく依存」は咎められるべきものではなく，むしろ合理的なものである．問題となるのは,「過信に基づく依存」である.

さて,「平時に作動するシステムに対して，過信（あるいは過信に基づく依存）は起こり得るだろうか」と問われたとすると，その答えは肯定的なものとなろう．過信の可能性が否定できない理由として，たとえばつぎの２つをあげることができる.

第１は,「平時に作動するシステム」であるがゆえに，ヒトはそのシステムが「知的に振る舞う」様子

を何度も繰り返して見ることができるという点である．システムの振る舞いを日常的に眺めているうちに，「A のような場面では，システムは B のように振舞う」といったように，ヒトは自分の体験を通じてシステムのメンタルモデルを構築していく．日常に遭遇するさまざまな場面のなかでシステムが見せる挙動に満足感を覚えるようになると，ヒトは「システムに任せて安心」という気持ちを抱くであろう．しかし，いかに日常的とはいえ，過去に遭遇した場面と寸分たがわぬものばかりが現れるわけではない．場面 A に似ているが，実はそれとは本質的に異なる A*が出現することもある．ヒトは，いつもの場面 A だと思ってシステムに任せるかもしれない．しかし，A*がシステムの能力を超えるものだったらどうだろう．「システムが対応してくれるはず」と思っていても，システムはヒトが期待したとおりの機能を発揮することはない．

第２の理由としては，「システムの挙動が自分の想像（期待）していたものとは違うといった事態が生じたとしても，人がそれに対応できるだけの十分な時間的なゆとりがある」という，「平時」が持っている本質的な特徴をあげることができる．

では，「緊急時に作動する（はずの）システム」に対する過信は発生するのだろうか．これについては，慎重に検討したうえで答えを出す必要があるが，過信が生じたとしても，「平時に作動するシステム」に対する過信ほどの頻度は持たないのではないかと考えられる．その理由は，少なくとも２つある．

第１は，「緊急時に作動するシステム」が本当に作動する場面にヒトが遭遇する機会は少ないことである．「緊急時に作動する」ように設計されているとは聞いている（あるいは，知っている）としても，自分の眼で見る（体験する）機会が少ないものに対して，ヒトは「任せて安心」と感じるだろうか．

第２は，もし万一，「システムの挙動が自分の想像（期待）していたものとは違うといった事態が生じたなら，もはや人がそれに対応できるだけの十分な時間的なゆとりはない」という，「緊急時」の本質的特徴があげられる．極論すれば，「自分の命をかけてまで，システムに頼ろうと考えるヒトはいるのか」と言い表すこともできる．

6. 機械がヒトの意向に逆らうことは許されるか

6-1 人間中心の自動化

ヒューマン・マシン・システムでは，「権限の所在」はつねに大きな議論の対象となる．例えば，航空分野における「人間中心の自動化」では，「人に権限を与えておく」ことが基本的な要請とされる．しかし，「いついかなる場合でも，人に最終的な権限を与えておく」設計が正しいとは限らない[7]．

さらに，航空以外の分野においても「人に権限を与えておく」ことが最重要であるとは限らない．自動車の衝突事故を解析してみると，有効な手段を何も講じないまま（すなわち，ステアリングでの回避も行わず，ブレーキもかけないまま）障害物にぶつかっていくケースも少なくないことが知られている．障害物への急接近が検知されたとき，システムは「ブレーキをかけよ」と警報を発するだけでよいのだろうか．ブレーキをかけるかどうかの最終判断はドライバが決めればよいという考え方は，「人に権限を与えておく」ものであるとはいえるが，本当にそれでよいのだろうか．自動車の分野では，ドライバの操作が遅れているときは，システムが自らの判断で自動的にブレーキをかける PCS が実用化されているが，これは，厳密な意味では旧来の「人間中心の自動化」の枠を超えたシステムである．

6-2 安全確保の視点から権限を考察すると

ヒトと機械の間での役割分担を定めることを「機能配分」[8]とよぶ．ヒューマン・マシン・システム設計における重要な課題のひとつである．機能配分にはさまざまな方式があるが，代表的なものが「考察の対象となる機能ごとに人間と機械の能力を比較し，優れた能力を持つほうに当該機能を割当てる」といった方式である．しかし，個々にはヒト向きのタスクであっても，そのようなタスクを複数個同時に担当しなければならない状況では，必ずしもヒトの優位性は保証できない．さらに，タスクの担当が長時間に及ぶときは，さまざまな要因によってヒトの作業の質や効率は低下することがある．ただし，しばらくすると，ヒトは本来の優れた能力を取り戻すことも多い．このことから，ヒトが担当していたタスクを機械に移す，あるいはその逆向きの受け渡しを行うことが必要となる．一方の主体が行っていたタスクを，ある時点で他方の主体に譲り渡すことを，そのタスクに関する「権限委譲」[7]という．

では，ヒトと機械が担当している各タスクについて，「当該タスクの権限委譲を行う必要はあるか．行うとすれば，それはいつか」を決める権限は，いったい誰が持つのだろうか．人でなければならないのだろうか，それともコンピュータであってもよいのだろうか． この問題を「システムの安全確保」の視

点から考察してみよう．

あるヒューマン・マシン・システムにおいて，ヒトが直面している「状況」をひとつ想定しよう．その状況の中でのヒトの行為は，つぎの３種類のうちのいずれかひとつである．（H１）必ず実行しなければならないもの，（H２）行っても，行わなくても，いずれでもよいもの，（H３）決して行ってはいけないもの．また，「コンピュータがヒトをモニタできるようになっている」ものとしよう．ヒトの行為に対するコンピュータの判断は，つぎの２通りのうちのいずれかである．（C１）ヒトの行為が検出された，（C２）ヒトの行為は検出されていない．ここに述べた状況を図示したものが図２である [2)].

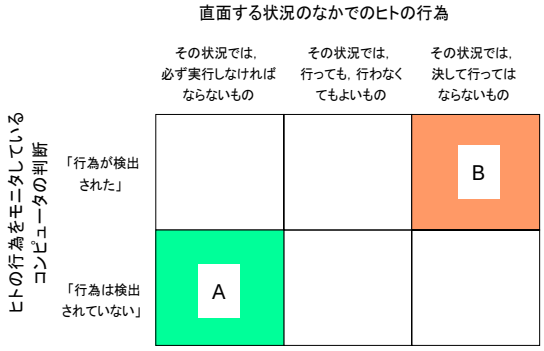

図２　機械の判断による権限委譲

図２における領域 A は，「直面している状況の中で，ヒトはなすべきことをしていない」とコンピュータが判断するケースを示す．先行車が急減速しているので，直ちにブレーキをかけなければならないのだが，ドライバのブレーキ操作がまだ検出されないといったケースである．ドライバは脇見をしていて，先行車の急減速に気づいていないのかもしれない．あるいは，先行車が急減速するのを見てパニック状態になり，ブレーキ操作ができないのかもしれない．このようなとき，ドライバに指示されていないからといって，コンピュータはブレーキもかけず傍観していてよいのだろうか．それとも，自らの判断で緊急ブレーキをかける（ブレーキ操作に関する権限を，ヒトからコンピュータに委譲させる）ことが適当なのだろうか．ここで，機械が自律的に安全制御を行うことは，「ドライバの操作の欠落部を機械が補う」ことであることに注意すれば，そのような場面で「機械に権限を与える」ことには，さほどの違和感は覚えないものと思われる．ただし，それは，「本来必要な操作をドライバが行っていない」との機械の判断が正しい場合に限られる．

図２における領域 B は，「直面している状況の中で，決してやってはいけないことをヒトがしている」とコンピュータが判断するケースである．隣のレーンを背後から高速で接近する車があるのに，ドライバが車線変更をしようとしていることが検出されるといったケースである．現在の自動車技術では，ドライバのステアリングにはかかわりなく，車輪の向きを制御することが可能であるが，「不適切な車線変更」が行われようとしたとき，コンピュータがそれを阻止する（ステアリングに関する権限を，ヒトからコンピュータに委譲させる）ことは許されるだろうか．このような場面で，「ドライバの行為を機械が完全に抑制・阻止する」ハード・プロテクションの考えかたを採るべきか，「あくまでもドライバは何らかの明確な意図を持って行おうとしているのであろうから，本来はその状況にそぐわない行為であることのみをドライバに知らせればよく，その注意喚起をオーバーライドしてドライバがその行為を続けることは認める」とのソフト・プロテクションの考え方を採るべきかは，議論が分かれる．

参考文献

1) Endsley, M.: Towards a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1), 32-64, 1995.

2) 稲垣: 自動化による安全性の向上：ヒューマンファクターの視点からの考察，Fundamentals Review, Vol. 2, No. 2, pp. 20-30, 2008.

3) 稲垣: リスクを見つけて制御する，安全飛行（全日空），No. 244, pp. 2-14, 2006.

4) 稲垣: 人が機械を知り，機械が人を知る，第 16 回交通・物流部門大会講演論文集, pp. 7-10, 2008.

5) Lee, J. & Moray, N.: Trust, control strategies and allocation of function in human machine systems; *Ergonomics*, 35(10), 1243-1270, 1992.

6) 稲垣:　人間機械共生系：システム設計の視点と課題，自技会「ヒューマトロニクス」シンポジウム，2005.

7) 稲垣:“リスク環境における人と知能機械の協調をデザインする，”電子情報通信学会誌，vol. 89, no. 12, pp. 1026-1031, 2006.

8) 稲垣: 人間と機械の機能分担，自技会「人と技術の協調によるアクティブセイフティ」シンポジウム，2004.

（注）上記文献のうち，2) - 4), 6) - 8) については，http://www.css.tsukuba.ac.jp（筑波大学認知システムデザイン研究室）「解説記事」のページからダウンロード可．