CC-003

人工知能の制御へのより安全な適用について
A Study on Safer Employment of the Artificial Intelligence to Control

金川　信康†
Nobuyasu Kanekawa

1. はじめに

自動運転を初めとする制御の全自動化は，人為的操作を不用とし，人為的誤りに起因する事故の確率を低減し，安全性を向上させることが可能となる。一方，緊急時には人為的なオーバーライド（操作介入）を可能とすることで，制御装置の故障や想定外の事態が発生した場合の安全性を確保することができる。第２次 AI（人工知能）ブーム中に提案された「ニューラルネットワーク」を発展させた「ディープ・ラーニング」（深層学習）の研究が起爆剤となり，現在再び AI の研究が活発に進められており，第３次 AI ブームと呼ばれている[1]。ディープ・ラーニングを初めとする人工知能は人知を超えた最適解を提供してくれるが，その安全性は必ずしも保証されたものではない[2]。そこで人工知能の動作の安全性を検証，保証する技術を付加することにより，人知を超えた安心・安全な最適解を得られることが期待される。

2. 安全検証型適応制御

2.1 知能化制御と安全性

深層学習に代表される知能化機能は人知を超えた最適解が期待できるが，内部状態が不明であるか，たとえ情報を得ることができても人間が理解できる形での表現は困難で，そのままでは安全性は保証されていない。そこで知能化制御 (Intelligent Control)に安全検証機能 (Safety Verification)を付加することにより人知を超えた安全な最適解を得られることが期待できる(図１)。

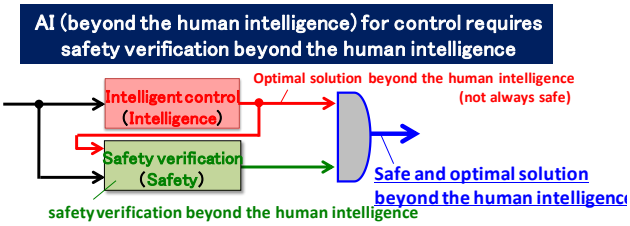

図１　安全検証の必要性

自動運転にかかわる過去の事故事例を分析すると，日照条件による画像認識誤り，人間―機械双方による「だろう運転」，不適切な人為的介入（オーバーライド）が主な原因としてあげられる。

これらの内自動運転システムに起因する要因，すなわち，日照条件による画像認識誤り，機械による「だろう運転」は先にあげた，知能化制御に対する安全検証機能(Safety Verification)により回避可能で，さらに入力データ補完により，機械側に起因する危険事象を回避可能である(図２)。

† （株）日立製作所, Hitachi, Ltd.

さらに人為的要因，すなわち，人間による「だろう運転」，不適切な人為的介入（オーバーライド）も人為的操作に対する安全検証機能により回避可能で，さらに安心志向制御により人間に不必要な不安感を抱かせないようにでき，不適切な人為的介入（オーバーライド）の根本原因となる不必要な人為的介入の機会を減らすことができる。

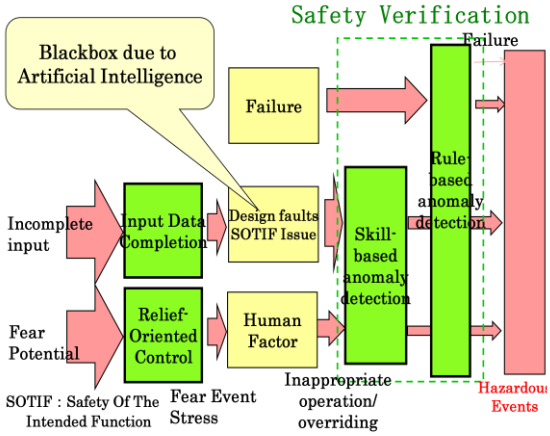

図２危険事象発生メカニズム

そこで，知能化制御[3]，人為的操作に対する安全検証機能[4]，入力データ補完[5]，安心志向制御を備えた安全検証型適応制御(図３)を提案する。

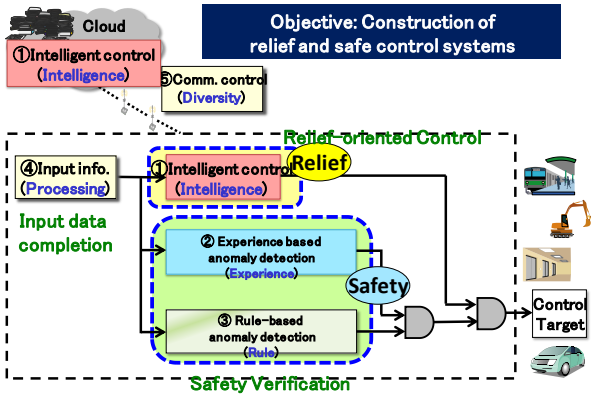

図３　安全検証型適応制御

2.2 技術課題と対応策

安全検証型適応制御の当初案では安全第一の思想の下，高性能知能化制御(High-performance Intelligent Control，以下「知能化制御」と称す)の出力を AND ゲートにより抑制してフェールセーフ性を確保するアーキテクチャであった。しかし，自動運転などの多くの応用分野では，フェールセ

ーフ性に加えてフェイルオペレーショナル性（故障時動作継続性）が求められている。

そこで，フェイルオペレーショナル性（故障時動作継続性）を実現するために以下の方策を講じる。

・リミッタ型出力制限方式
・オペレータオーバライド
（優先度付／力覚フィードバック付）
・知能化制御の冗長化（ダイバーシティ）

2.3 リミッタ型出力制限方式

AND 型出力制限方式の構成を図４に示す。安全検証機能には入力および知能化制御出力が入力され，それらに対応する検証出力(OK/NG)が出力される。さらに過去の値からの状態遷移にも着目した（遷移チェック付の）場合には，１サンプル前（z^-1）の過去の入力および知能化制御出力も入力され，それらに対応する検証出力(OK/NG)が出力される。AND 論理は検証出力が OK のときにのみ知能化制御出力を検証済出力(Verified output)として出力し，NG の時には出力を停止する。

安全検証機能の動作は図４の下側に示すように入力および知能化制御出力，さらには遷移チェック付の場合には過去の入力および知能化制御出力の組み合わせをエントリーとして，それらに対応した検証出力(OK/NG)が出力される。

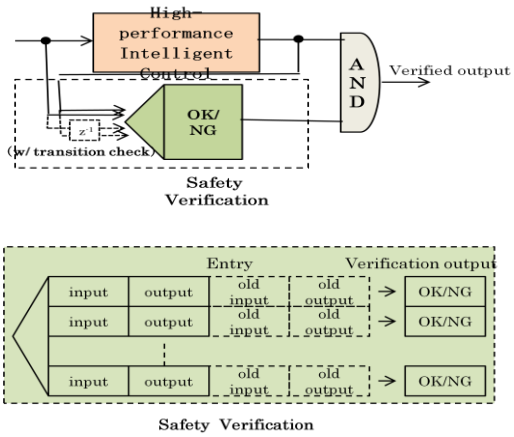

図４　AND 型出力制限方式

リミッタ型出力制限方式の構成例を図５に示す。安全検証機能には入力および知能化制御出力が入力され，それらに対応する上限及び下限が出力される。さらに過去の値からの状態遷移にも着目した（遷移チェック付の）場合には，１サンプル前（z^{-1}）の過去の入力および知能化制御出力も入力され，それらに対応する上限及び制御出力下限が出力される。

制限値選択回路では，検証済出力として入力された知能化制御出力が上限と下限の間にある場合には知能化制御出力を検証済出力として出力し，入力された知能化制御出力が上限を上回る場合には上限に制限した値を検証済出力として出力し，入力された知能化制御出力が下限を下回る場合には制御出力下限に制限した値を出力する。さらに，制限値選択回路または安全検証機能，知能化制御出力は安全検証結果をステータスとして出力する。ステータスは OK（制御出力下限～上限の範囲内である），OK w/limit（下限～上限の範囲外であるが，下限～上限の間の値が存在する。すなわち，下限＜上限が成り立つ。），NG（下限～上限の間の値が存在しない。すなわち，下限＜上限が成り立たない。）の３値とする。

同様に，安全検証機能の動作は図５の下側に示すように入力および知能化制御出力，さらには遷移チェック付の場合には過去の入力および知能化制御出力の組み合わせをエントリーとして，それらに対応した上限及び下限が出力される。

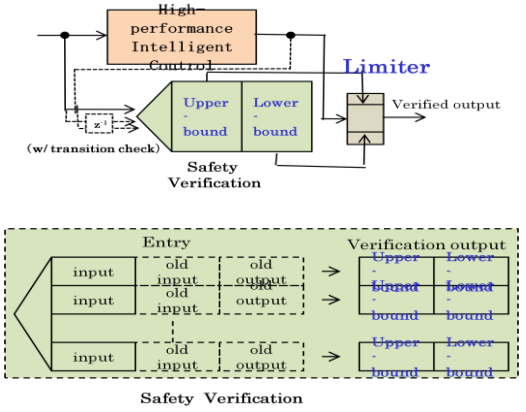

図５　リミッタ型出力制限方式

以上，簡単のために１変数制御について述べたが，実際の制御対象では多変数制御である場合が多い。自動運転を例に取ると，加減速度，ヨーレートの２変数を制御することが考えられる。以下，多変数制御，特に簡単のために２変数制御の例について説明する。

図６は知能化制御がα，βの２変数を出力する場合の例である。変数αのための安全検証機能は入力および知能化制御出力のβ成分の組み合わせをエントリーとして，それらに対応したα成分の上限及び下限が検証済出力として出力される。変数βのための安全検証機能は入力および知能化制御出力のα成分の組み合わせをエントリーとして，それらに対応したβ成分の上限及び下限が検証済出力として出力される。

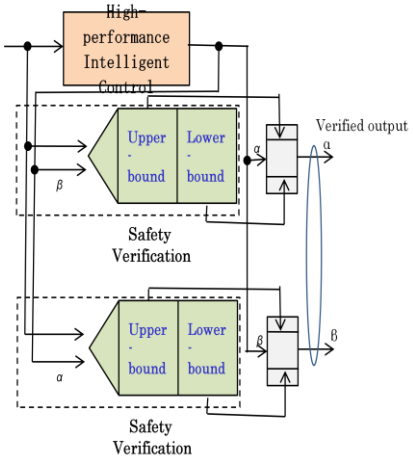

図６　多変数制御

なお，変数が 3 以上に場合においても同様に実現が可能で，一般化すれば，変数 x_i のための安全検証機能は入力および知能化制御出力の他の全ての成分 x_j $(j \neq i)$の組み合わせをエントリーとして，それらに対応した x_j 成分の上限及び下限が検証済出力として出力される

図 7 は知能化制御出力（α_i，β_i）の制限方法の具体例である。

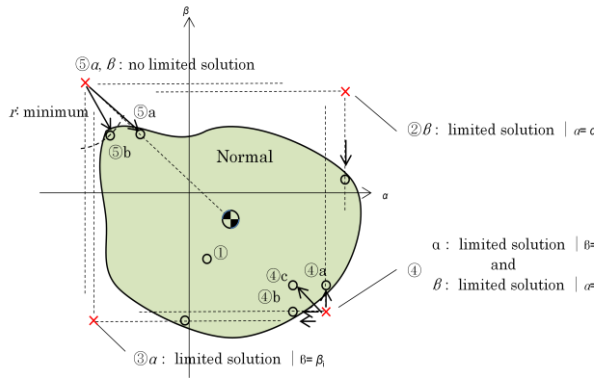

図 7　多変数制御における制御値制限

①は α，β 成分ともの上限及び下限の範囲内の場合で，知能化制御出力が制限を受けずに検証済出力として出力する。

②は $\alpha = \alpha_i$ において，β に制限解がある場合で，β の値が制限されて検証済出力として出力する。

③は $\beta = \beta_i$ において，α に制限解がある場合で，α の値が制限されて検証済出力として出力する。

④は α，β ともに制限解がある場合で，

④a：β の値が制限されて検証済出力として出力される，

④b：α の値が制限されて検証済出力として出力される，

④c：α，β ともに値が制限されて検証済出力として出力する

のいずれかが考えられる。

⑤は α，β ともに制限解がない場合で，

⑤：出力停止，

⑤a：正常領域の重心方向へ修正した点の α，β を検証済出力として出力する，

⑤b：正常領域の最短距離へ修正した点の α，β を検証済出力として出力する

のいずれかが考えられる。

ここで，④a～④c のどの方法を選ぶかは，α，β のどちらの値の制限を優先させるべきかに依存し，アプリケーションまたは入力 1 として与えられる状況に依存する。自動運転を例に取ると，制御目標値として，速度とヨーレートが考えられ，両者は遠心力がタイヤのグリップ力を上回らないように制限される。進行方向や横方向に障害物がある場合には，ヨーレートは制限せずに速度を制限し，横方向に障害物がない場合にはエネルギ効率を高める（原則によるエネルギ損失を減らす）ために，速度よりもヨーレートを制限する。

④a は β の値を制限させることを優先させる場合で，④b は α の値を制限させることを優先させる場合で，④c は α，β の両方の値を制限させなければならない場合である。

さらに⑤は最も実現が簡単な例で，⑤a，⑤b は入力，α 出力，β 出力をエントリーとする安全検証機能とする必要がある。

2.4 ペレータオーバライド

(1)　優先度付オペレータオーバライド

優先度方式オペレータオーバライド方式を図 8 に示す。制御システムは図のように操作インタフェースに加えて，自動制御機能，安全検証機能を有する。操作インタフェースには操作者により操作量の物理量が入力され，操作量情報と優先度情報に変換される。自動制御機能は入力に基づき制御出力を出力し，優先度情報が 1 よりも小さいときに，切り替えスイッチ SW1，リミッタ，切り替えスイッチ SW2 を経て制御出力を出力して制御対象を制御する。

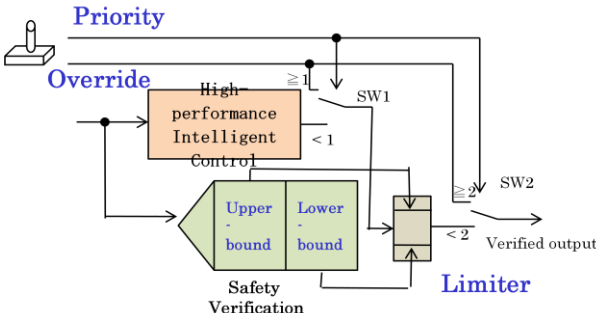

図 8　オペレータオーバライド

安全検証機能は入力と切り替えスイッチ SW1 の出力から安全性を検証し，安全性が確認された場合にはリミッタはスイッチ SW1 からの入力を出力し，安全性に問題がある場合には切り替えスイッチ SW1 の出力を遮断または出力を制限して，リミッタから出力される。

優先度情報が 1 よりも大きいときには，切り替えスイッチ SW1 は操作量情報を選択し，安全検証機能の制限下で制御出力となる。優先度情報が 2 よりも大きいときには，切り替えスイッチ SW2 は操作量情報を選択し，安全検証機能の制限を受けずに制御出力となる。

以上の動作によれば，安全検証機能により自動制御機能からの危険な出力を防止でき，安全な動作を保証できる。さらに，介入操作の優先度を入力する手段を設けることにより，緊急時の状況により，安全検証機能の制限化での緊急介入操作，安全検証機能の制限を受けない緊急介入操作の切り替えが可能となる。

自動制御機能として，深層学習，機械学習などの人工知能を導入することにより，人知を超えた制御性能を実現することが期待されるが，人知を超えるがゆえに，安全に関してのアカウンタビリティ（説明責任，説明性）を向上させることが望ましい。そこで本発明により，安全検証機能を付加することで，人工知能による人知を超えた高度な制御を安全に実現することが可能となる。

なお，切り替えスイッチ SW1，切り替えスイッチ SW2 で一方の入力から他方の入力に切り替える瞬間に両者の間に差があるため切り替え結果 out には両者の差に相当する段差が発生する。そこで，例えば時間をかけて両者の値を

徐々に切り替えることにより段差を発生させるに切り替えることが出来る。

さらに切り替えスイッチ SW1，切り替えスイッチ SW2 を加算機能 ADD1，ADD2 とすることにより，切り替えスイッチ SW1，切り替えスイッチ SW2 の切り替えに伴う段差なく制御出力を生成でき，より滑らかな制御が可能となる。

操作インタフェースの操作自由度を高め，本来の操作量の物理量とに割り当てられている操作自由度とは独立した操作自由度により優先度情報を生成する例を図９に示す。

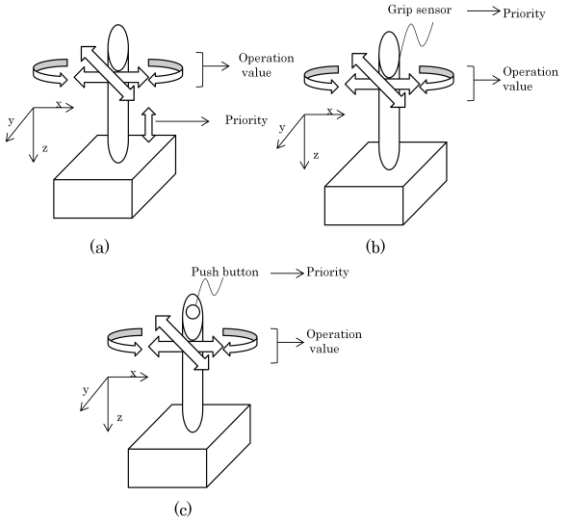

図９　操作インタフェース

図９(a)はｘ－ｙ平面内の動作を本来の操作量の物理量とし，ｚ軸方向の操作により優先度情報を生成する例である。(b)はｘ－ｙ平面内の動作を本来の操作量の物理量とし，握力センサにより計測された握力により優先度情報を生成する操作インタフェースの例である。(c)はｘ－ｙ平面内の動作を本来の操作量の物理量とし，ｘ－ｙ平面内の動作を本来の操作量の物理量とし，押しボタンの押し量または力により優先度情報を生成する操作インタフェースの例である。

以上図９に示した操作インタフェースの構成例によれば，本来の操作量の物理量とは独立して優先度情報を生成することが可能となる。

また，複数の操作インタフェースを備えておき，より大きな数の操作インタフェースに操作量の物理量が加えられた場合に，より高い優先度情報を生成する実施例も可能である。本方式に拠れば，複数の操作者により同一の操作量の物理量が加えられた場合に，より高い優先度情報を生成することがでできる。

(2)力覚フィードバック付オペレータオーバライド

図１０は安全検証機能の上限，下限に基づき操作インタフェースに力覚フィードバックを施す方式である。

インタフェースに加えられる操作量（物理量）と反力の関係は図１１の通りで，不感帯（オーバーライドしない操作領域）とオーバーライド領域との境界，上限，下限に相当する操作量（物理量）のところで，反力に不連続点を有する。

なお，優先度情報は操作量（物理量）が不感帯（オーバーライドしない操作領域）の場合には０，オーバーライド領域の時には１を出力する。

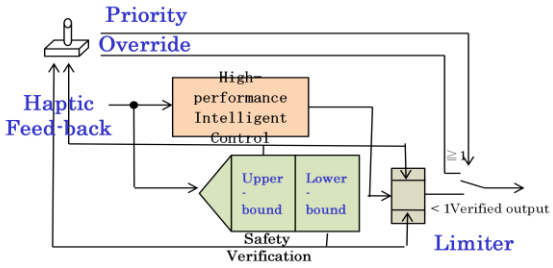

図１０　オペレータオーバライド

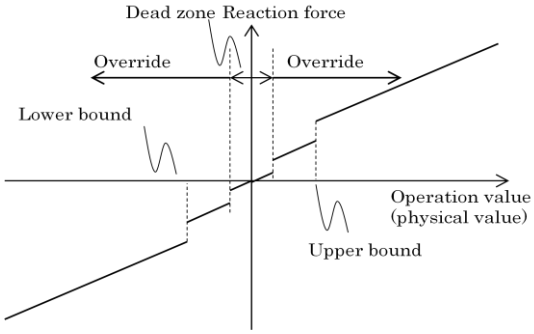

図１１　操作量と反力の関係

本方式に拠れば，操作者に不感帯とオーバーライドとの境界，上限，下限の境界に相当する操作量（物理量）を認識させることができ，オーバーライドしない操作領域とオーバーライドする操作領域の区別，上限，下限範囲内と範囲外の区別をつけることができ，それらの領域を意識した操作を可能とする。

2.5 システム全体構成

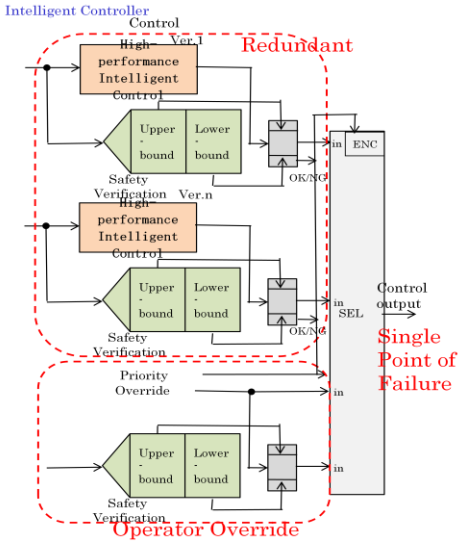

図１２　システム全体構成

図１２はオペレータオーバライドに加えて，自動制御機能を冗長に備えた論理的実装例である。出力選択機能では制限値選択回路からのステータスに基づき，制限値選択回路からの検証済出力，オーバーライドの操作量情報のなかから１つを選択して制御出力とする。

図１３はシステムの物理的実装の例である。図１２に示すシステム構成のままでは，出力選択機能が単一故障点（single point of failure）となり，出力選択機能の故障がシステム全体の故障につながり，システム高信頼化のボトルネックとなる。

そこで，図１３に示すように制御出力に従って動作するエッジ側の制御機能ごとに出力選択機能を設け，それぞれのエッジ側の制御機能と出力選択機能を個別の制御装置に実装することが考えられる。

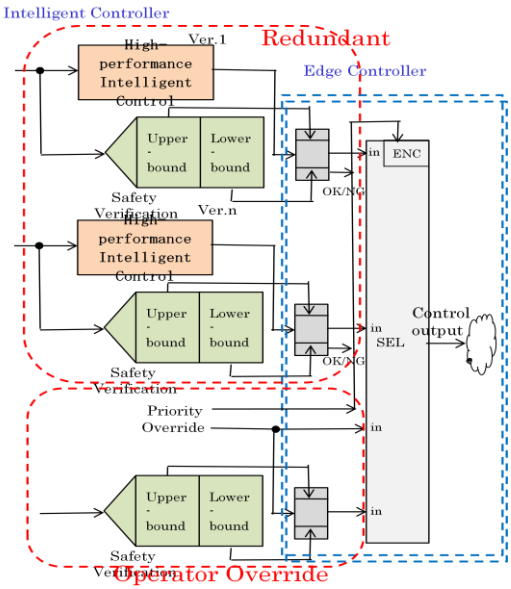

図１３　物理実装

2.6 信頼度評価

(1)人為的操作に対する安全検証機能，安心志向制御の効果

2.1 節において，人間による「だろう運転」，不適切な人為的介入（オーバーライド）も人為的操作に対する安全検証機能により回避可能であると述べたがその効果について検証する。

一般的に人間の誤り確率は THERP (Technique for Human Error Rate Prediction)によると 0.05～0.3%であるといわれている[6]。実験結果により，安全検証機能による異常検出率は 90%を超えており，安心志向制御を適用することにより 90%以上の人が不安を覚えることなく安心して自動制御に委ねることができるようになることが報告者らの実験でわかっている。以上の結果を踏まえると，不適切な操作，オーバーライドの 90%は安全検証機能により検出されて危険事象に至らない。従って，人間の誤りにより危険事象が発生する確率はその 1/10 の 0.005～0.03%となると見積もることができる。さらに，安心志向制御を導入することにより不適切なオーバーライドを引き起こす要因となる不必要な不安感を人間に与える可能性も 1/10 となることから，さらに 1/10 の 0.0005～0.003%となると見積もることができる。

また，自動車による死亡事故発生率は 500 x 10-7 [/hr.]と報告されており[7]，人間の誤り率を故障率に当てはめると 50,000[FIT]に相当し，同様に人為的操作に対する安全検証機能，安心志向制御を適応することにより残存故障率は 1/100 の 500[FIT]となると見積もることができる。

(2) 単一故障点の排除の効果

2.5 節で示したように，制御出力に従って動作するエッジ側の制御機能ごとに出力選択機能を設け，それぞれのエッジ側の制御機能と出力選択機能を個別の制御装置に実装することにより出力選択機能が単一故障点となることを回避することができる。

ここで，エッジ側の制御装置それぞれの故障率を 500[FIT]と仮定する。この値は合理的，すなわち，制御装置を構成する電子部品の現在の品質で十分に実現可能な値である。エッジ側の制御装置は個々の機能ごとに２冗長構成(1-out-of-2)で，A, B, C３つの全ての機能が動作可能である場合にのみ全体システムが動作可能であるとすると，システム全体の信頼度は図１４に示すような信頼性ダイアグラムにより表され，次式により算出できる。

$$R = \{ 1 - (1 - R_i)^k \} \{ 1 - (1 - R_e)^2 \}^3$$

$$R_i = exp\ (- \lambda_i\ t)$$

$$R_e = exp\ (- \lambda_e\ t)$$

但し，　R_i：知能化制御装置の信頼度
R_e：エッジ側制御装置の信頼度
k：（$=b+1$）知能化制御装置の冗長度
b：　知能化制御装置のバックアップ数
λ_i：知能化制御装置の故障率（1000[FIT]）
λ_e：エッジ側制御装置の故障率（500[FIT]）
t：経過時間

ただしここでは各ノードはセルフチェッキングであり，誤り検出カバレッジは簡単のために 100%であると仮定する。

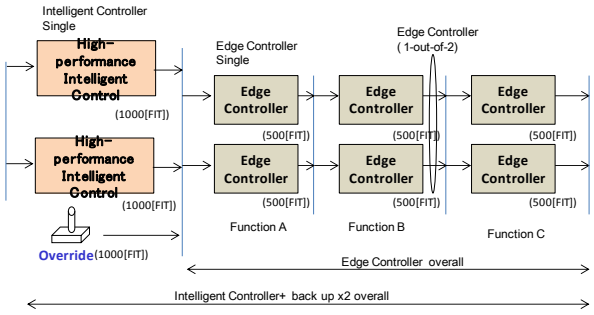

図１４　信頼性ダイアグラム

上記仮定に基づきエッジ制御装置の信頼度を求めると図１５のようになる　100000 時間後の信頼度は図１５に示す

ように，エッジ側の制御装置が単一である場合には 0.951 であるのに対して，エッジ側の制御機能と出力選択機能を個別の制御装置に実装することにより，エッジ側制御装置(1-out-of-2)が 0.998，機能 A, B, C を全て合わせたエッジ側制御装置全体が 0.993 と大幅に向上し，SIL (Safety Integrated Level) 3 を実現するための残存故障率(100[FIT])よりも高い信頼度となっていることがわかる。

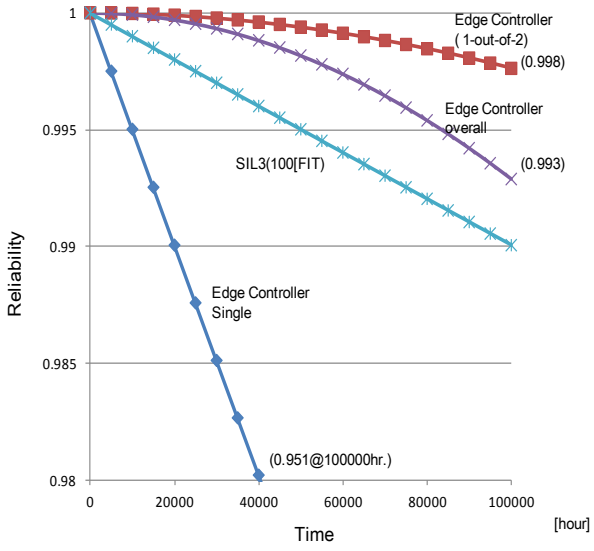

図１５　エッジ制御装置の信頼度

(3)知能化制御の冗長化の効果

続いて，知能化制御装置の冗長化の効果を検討する。知能化制御装置はエッジ制御装置と比べ，性能，機能も高いためより複雑で故障率もより高いと考えるのが妥当である。ここで知能化制御装置の故障率を 1000[FIT]と仮定する。現時点では開発の緒についたばかりで部品点数も多く，この仮定よりも 1 桁程度大きいと考えるのが妥当であるが，今後の集積化を考慮すると合理的，すなわち数年後には十分に実現可能な値である。

知能化制御装置単体の信頼度は図１６に示すように 100,000 時間後に 0.905 と非常に低いものとなってしまうが，知能化制御装置にバックアップ機能を設けると信頼度は向上する。ここでバックアップ機能の故障率を知能化制御装置と同じく 1000[FIT]と仮定すると，100,000 時間後の信頼度は 0.991，エッジ側制御装置を含めたシステム全体で 0.984 となる。ここで，バックアップ機能は冗長に設けた知能化制御装置が第一に考えられ，次にオペレータによるオーバーライドが考えられる。先に述べたように人為的操作に対する安全検証機能，安心志向制御を適応することにより残存故障率は 1/100 の 500[FIT]となると見積もられることから，オペレータによるオーバーライドは知能化制御装置異常時のバックアップ手段としての役割を果たすことが可能であるとみなせる。

さらにバックアップ機能を２つ設けることにより，エッジ側制御装置全体を含めたシステム全体で信頼度は 0.992 となり，SIL3 を実現するための残存故障率(100[FIT])よりも高い信頼度となる。ここでの２つのバックアップ機能は先に述べたように冗長に設けた知能化制御装置が第一に考えられ，次にオペレータによるオーバーライドが考えられる。

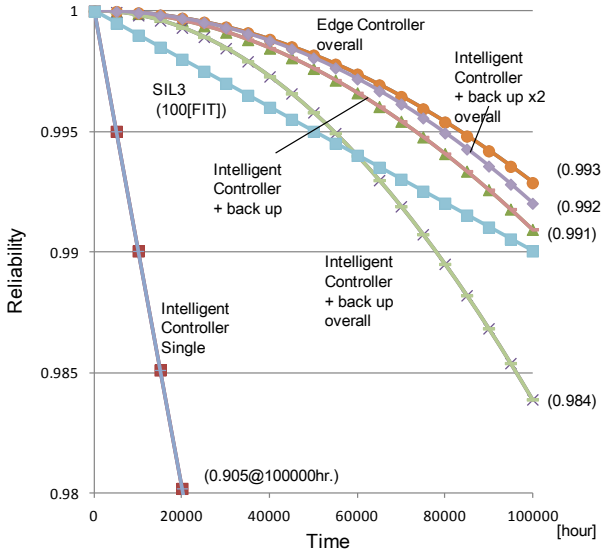

図１６　システム全体の信頼度

3. おわりに

知能化制御に安全検証機能を付加し，人為的操作に対する安全検証機能，安心志向制御の導入に加えて，新たにフェイルオペレーショナルに最適なリミッタ型出力制限方式，出力選択機能が単一故障点とならないシステム構成（実装方式）を提案し，フェイルオペレーショナルを実現できる目処を得た。

謝辞

本研究の機会を与えていただいた（株）日立製作所研究開発グループ各位、知能システム制御ラボラトリ各位に心より感謝いたします。

参考文献

[1] 人工知能の歴史 - Wikipedia
http://ja.wikipedia.org/wiki/人工知能の歴史.
[2] Stephen Hawking warns artificial intelligence could end mankind
http://www.bbc.com/news/technology-30290540.
[3] 広津他, 深層学習を活用した高精度知能化制御の提案, FIT2017, CF-007 (2017)
[4] 中川,組み込みシステム向け異常検知方式, FIT2017, F-013 (2017)
[5] 西田, 奥出, 隠れマルコフモデルを用いた複数個体による高信頼環境情報の推定技術, FIT2017, CO-014 (2017)
[6] A. D. Swain, H. E. Guttmann: Handbook of human reliability analysis with emphasis on nuclear power plant applications, NUREG/CR-1278 (1983)
[7] 林喜男　システム安全，安全工学 18, 6 (1976)