The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

将来の機械知性に関するシナリオと分岐点

Scenarios and branch points to future machine intelligence

高橋恒一[*12]
Koichi Takahashi

[*1] 理化学研究所　生命機能科学研究センター
RIKEN Center for Biosystems Dynamics Research

[*2] 慶應義塾大学大学院政策・メディア研究科
Keio University Graduate School of Media and Governance

We discuss scenarios and branch points to four major possible consequences regarding future machine intelligence; 1) the Singleton scenario where the first and single super-intelligence acquires a decisive strategic advantage, 2) the Multipolar scenario where the singleton scenario is not technically denied but political power-game or other factors in human society or game theory between intelligent agents inhibit a single agent to acquire a decisive strategic advantage, 3) the Ecological scenario where the singleton scenario is technically denied and many intelligent autonomous agents operate in the manner where they are mutually dependent and practically unstoppable, and 4) the Upper-bound scenario where cognitive abilities that can be achieved by human-designed intelligent agents or their descendants are inherently limited to the sub-human level.

1. はじめに

本稿では，機械知性の発達を比較的遠距離の将来までに渡るシナリオとして展開した場合にありうる結果をいくつかに分類し，それぞれに至る道程に存在すると想定される主要な分岐点を見いだすことを試みる．

次のような前提を設定し，議論を主要な論点に絞る．第一に，物理法則により禁止されていない技術は，十分な資源の投入を条件にいつか実現される．第二に，物理的に禁止されない技術のうち経済的に合理的な資源投入の結果それに見合う経済的あるいはその他の効用が予見される技術開発は，いつか何らかの主体によって成される．ただし，要素技術と結合技術には相対的な関係性が存在し，技術間には依存関係のグラフが存在する．第三に，技術拡散は一定の速度で起き，発明された技術はある時間の後に共有される．第四に，過渡的な力の不均衡は，一定の時間の後に均衡が示す安定点に収束する．ただし，過渡的動態のうちに存在する不可逆的な分岐点は不確実性要因となる（偶然の固定）．

なお，ここでは高度な知性を備える知能エージェントを機械知性と呼ぶが，これは現在の人工知能という用語の使われ方の多様さによる混乱を軽減することのみが目的である．また，本稿は特定の年代を想定した未来の予測を目的とするものではなく，道程と分岐点およびそれらのありうる帰結を整理する試みである．

2. 機械知性シナリオの分類

ここでは，長期的な機械知性技術の発展により想定されるシナリオを整理する．

2.1 シングルトンシナリオ

再帰的に自己更新する知能エージェントの自己改良速度は上限なく，あるいは，その時点でその改良速度を観察する人類またはその他の知能エージェントに対する決定的戦略的優位性[Bostrom14]の獲得まであたかも上限がないように増大し，最初に自己更新を一定程度まで進めた知能エージェントが資源獲得や他のエージェントとの競争において覆すことが困難な覇権を握るとするシナリオである[Bostrom14].

連絡先：高橋恒一，理化学研究所生命機能科学研究センター，大阪府吹田市古江台 6-2-3，ktakahashi@riken.jp

2.2 多極シナリオ

シングルトンによる覇権の危険性の認識による国際条約の設定や技術経済的要因，あるいはその他の外部的な要因により，どのエージェントも決定的戦略的優位性の獲得する前に性能向上が停滞するが，関係主体間の力関係の変化あるいはテロリズムなどの不確実性要因によるシングルトンの発生の可能性は否定されないシナリオである[Bostrom14].

2.3 生態系シナリオ

知能エージェントの性能向上には決定的戦略的優位性の到達可能性以前に限界が存在し，その結果多数のエージェントが並存的，相互依存的な生態系様のマルチエージェントネットワークを構成するとするシナリオである．そのような「AI 生態系」は，今日のインターネットや電力網がそうであるように人類の活動が依存するため停止不可能であり，また個々のエージェントの挙動がある程度以上複雑であれば全体としての挙動も予測不可能であると想定される[Yamakawa17].

2.4 上限シナリオ

人類が工学的に作り出しうる知能エージェントおよびそれを起点にした能力発展には一定の上限が存在し，将来にわたっても自ら判断して人による逐一の指示なしに自律的に動作する能力は獲得しないというシナリオである[西垣 16，井上 17].

3. 制約とシナリオ分岐

本稿で設定した前提条件において，機械知性シナリオのうちどれが具現化するかは，知能エージェントの内部構造および既獲得資源と推論性能に関わる内部的制約と，知能エージェントの制御の対象外である物理法則や，他の主体との関係性により生じるゲーム理論的構造などの外部的制約がある．

3.1 内部的制約

(1) 高度な自律性に関わる制約

自律性は例外的状況や環境変化などの幅広い状況に対して対処できるタスク汎用性と深く関連する．上限シナリオに分類さ

れる議論の多くでは，工学的に設計可能な認知アーキテクチャの認知能力がそのタスク汎用性においてヒト並みに到達するには克服不可能な障害が存在すると主張される．しかしながら，ヒトの脳内の神経結合と計算機能的に同一のシステムは，認知機能的にも同一の能力を持つとする唯物論的な立場に立てば，これは物理的に禁止されない．ただし，十分な性能を発揮する内部構造を設計するためにどのような周辺技術や要素技術が必要であり，どの程度の研究開発コストを費やす必要があるかは現時点では検証されていない．また，環境および身体との相互作用で生じる創発的な動態や学習で獲得する認知機能がタスク汎用性にどの程度の寄与があるかの定量的な議論もまだ十分になされていない．

(2) 自己構造改良能力に関わる制約

同様の論点として，ある能力を持ったエージェントが，仮に利用可能な計算能力を一定とした場合に，自己または外部の同等の能力を持ったエージェントの構造情報を起点に，それを上回る応答速度あるいは認知能力を持つエージェントの新規の構造情報を生成できるかどうか，すなわち自己改良能力に関わる制約がある．これは一見遠大な議論に思えるが，実際には，上記の内部構造と認知能力に関わる制約が一定程度クリアされるならば，本質的には内部構造のモジュール性が関わる工学的な問題である．モジュール性と階層性（準分解可能性[Simon96]）が確保されたアーキテクチャであれば，設計主体の認知能力に合わせて様々な階層におけるモジュールごとに内部構造を改良する部分改良問題を一つずつ解いてゆけば，結果として漸進的に全体の継続的な能力向上が可能である．ただし，ヒトの脳を含め，ヒト並みあるいはそれ以上の能力を持った知能エージェントを準分解可能システムとして構築可能であるという仮説は未検証である．また，改良速度の上限は設計の変更による性能向上を予測する推論サイクルの実行速度により定義される．遺伝的アルゴリズムによる進化計算的アプローチも可能であるが，これもシミュレーションに必要な時間が速度の上限を定義する．

(3) 熱力学的効率に関わる制約

利用できるエネルギーを一定とした場合に実現可能な計算量には熱力学的な上限が存在する．情報の消去などの論理的に非可逆な操作は1ビットあたりボルツマン定数 k と絶対温度 T を用いて kT ln 2 に相当する熱力学的エントロピーの上昇を伴うことが熱力学第二法則から導かれる（ランダウアー限界）．これは300K の常温でおよそ 2.87×10^{-21}J である．現代の計算機のスイッチングエネルギーがおよそ 10^{-17}J 程度であり[Theis17]，あと1万倍程度でこの限界に到達する．ヒトの脳と比較すると，10W 程度のエネルギーで 10PFLOPS 程度の処理を行ない，1FLOPS あたり仮に 10^3 から 10^4 回程度の非可逆ビット操作が起きるとするとランダウアー限界とは2，3桁程度の熱効率の開きしかない．以上の数字は大雑把な見積もりではあるが，ヒトの脳と同等の 10PFLOPS の計算処理は今後もコモディティー化が進むと考えられる一方，ヒトを大幅に超えたタスク汎用性や予測能力を現在の延長上のデジタル計算機で実現するためには大量の電力が必要であることを示している．純粋な観測により量子状態が破壊されないような種類の量子計算や DNA 計算のような情報が失われない分子機械を用いた可逆的過程を用いる場合にはこの限界は局所的には適用されないが，これらの実用化が他の技術との比較で相対的にどの程度の速さで進むかは現時点でははっきり予測出来ない．

3.2 外部的制約

ここまで内部的制約を整理したが，これらは実際には物理法則や利用可能な資源量のような外部的制約を変数とするものが多い．関連する制約を整理する．

(1) 物理素子による制約

あらゆる物理現象には時定数が付随する．あるエージェントが利用資源量を変化させずに能力や応答速度を大幅に向上させるためには，アーキテクチャが不十分にしか最適化されていない場合を除いて自己の物理的基盤の更新が必要である．新しい物理現象やその組み合わせを利用する場合には，それを利用した場合の性能向上に関する仮説生成と物理実験による検証が必要であるが，必要な時間は対象の物理時定数に依存する．知識発見にシミュレーションを用いる場合にも，新たな物理現象が関わらない創発的な原理に関するものであれば物理時定数よりも高速に実行出来る場合もあるが，新しい物理現象の探索の場合には，対象の物理現象が属するより階層の低く粒度の細かい計算を要するため，一般的には実時間よりも長い時間がかかる．実在論的仮説生成[渡部16]により未知の過程を仮定して同じ階層でのモデル検証を通じて探索する場合にも，多数の可能な仮説を物理あるいは仮想実験により検証する必要があり，より長い時間がかかる．これらは，いずれにしても自己改良速度を制限する要因となる．

(2) 相対的優位性に関する制約

時間定数に関する制約は，エージェント間の優位性に関して追加の前提条件を与える．多くのエージェントが相互作用するマルチエージェント的状況であり，かつ環境および他のエージェントに関して不完全な情報しか得られない状態において，他のエージェントよりも優位な地位，すなわち他エージェントの行動およびその帰結をより早く予測しそれに対処出来る状態，を獲得し維持するためには，他の全てのエージェントおよび環境を含めた予測の対象となる系の状態変化の予測能力が有意に優っている必要がある．ここで，有意に優っているとは，予測に用いたフレームに対して外部からの撹乱に対する予測対象系の応答時間，および対象とするエージェントとの物理的あるいはネットワーク的距離により規定される相互作用の遅延時間のどちらよりも十分に長い時スケールの予測を，それらの時スケールに比べて短い時間で行える能力を保持しているということである．

相対的優位の確立には，相互作用の時定数が関連することを述べた．このことは，二つの重要な帰結を導く．第一に，物理世界でのマルチエージェント状況において，他のエージェントに対して予測能力と応答速度が少しでも優っていることは直ちに行動上の優位性の十分条件とはならず，前段落で定義した意味で「有意に」優っている必要がある．さらに，計算複雑性の観点からは，多くの問題は問題を構成する要素の数 N に対して線形の計算量で解ける状況はむしろ特殊であり，多項式的あるいは指数的な計算量を必要とするため，一般に計算能力の増加に対して対数的な効用しか得られない．しかしながら，第二に，確率的条件により偶然得られた資源的優位がたまたま決定的な差別化に結びつき，その状況が他のエージェントが追いつき均衡状態に戻るまでにかかる緩和時間の間に当該エージェントが決定的戦略的優位性にまで結びつく技術的あるいは獲得資源的進展を得れば，それが結果的に固定化された偶然[de Duve05]としてシナリオ分岐を起こす可能性がある（確率ゆらぎによる決定的優位性シナリオ）．

(3) 局在性に関わる制約

エージェントの物理空間における自他の境界は，エージェントの制御下にあるセンサー系と作用系の配置により決定する．また，光速を一定とするとエージェントを構成する主要な計算素子間の空間距離がエージェントの応答速度の上限を規定する．（これは，ヒトの脳の物理的大きさと神経伝達速度が脳の 100ms 程度の認知的応答速度を規定しているのと似ている．） 情報統合理論によると，意識体験が別個の部分の集まりではなく，統一されたひとつの全体として体験されるためには，要素間に相互情報量で計られる情報的結合が必要であるとされる[Tononi16]．エージェントが空間的広がりを持つ場合，このような情報統合により得られた経験に基づいて判断を行い，応答する場合の速度は，要素間の通信に伴う光速に起因する遅延により上限が決定される．センサー系や作用系との通信が一方向である場合には空間距離はエージェント内部の応答速度を直接は制限しないが，環境や他のエージェントに関する情報取得や作用の遅延という形で結果的に応答速度を制限する．このような制限は，前項の相対的優位性に関わる応答速度の制限を通じて，エージェントの能力に関する空間的広がりの局在性に関わる要求，さらに間接的には利用できる分散計算資源量の上限も設定する．

エージェントは，将来起きる状況の変化の予測に基づいて自己のコピーあるいは自己の影響下にある他種のエージェントを任意の数あらかじめ遠隔地に配置しそれらと通信することにより，局所的な出来事に対する応答時間を低減すると同時に，推論および探索の分担をすることで，空間的局在性に関する制限の回避を試みることができる．しかし，複数のエージェントが全体として統一の経験情報に基づいた意思決定プロセスを共有するには，情報のコピーを伴う通信が必要である．分散システムにおけるブリュワーの CAP 定理は，ノード間の情報複製に関して，一貫性，可用性，分断耐性の3つの保証にはトレードオフがあり，一般的にはこれらの2つしか同時には保証出来ないことを示している[Lynch02]．このトレードオフは実際には絶対的なものではなく，障害復旧の遅延時間が変数としてはたらく（無限の時間待てるのであれば，全ての障害を復旧して一貫性，可用性，分断耐性を確保出来る）．このことと光速の限界を合わせて考えると，単一エージェントの場合だけでなく，複数のエージェントの集合の場合にも，相対的優位性の確保を目的とした場合に，ある一定の応答時間内において利用できる計算資源の空間的局在性の要求，およびそれにより生じる計算能力の上限に関わる制限が設定されるため，複数のインスタンスを生成することで得られた追加の計算資源を用いることで際限なく相対的優位性を追求できるわけではない．

4. シナリオと分岐点

前章で議論した制約を用いて，2章で分類した各シナリオへの分岐点を検討する．冒頭で述べたように，本稿は特定の年代を想定した未来予測が目的ではなく，ここではシナリオの分岐点の特定と整理のみを行う．

高度な自律性に関わる制約は，機械知性が将来ヒト並み以上の認知能力を獲得するということを物理的には禁止しないものの，現実的なコストでそのような技術開発可能かどうかは今後の検証次第であり，現時点では不明である．もう一つの可能性は，ヒト並みの認知性能は人手で設計した機械では発揮出来ないが，その代わりに**自己構造改良能力**を持った機械の設計は可能であり，結果としてヒト並み程度以上の機械知性を手にするシナリオである．これらの二つの道筋により**上限シナリオ**とその他のシナリオとの分岐が生じる．

シングルトンシナリオは，前章で述べた全ての制限が突破あるいは回避され，一つのエージェントが決定的戦略的優位性を確保するシナリオである．ここでの決定的要因は，決定的戦略的優位性の確保に必要な認知能力のレベルが何を基準に設定されるかである．単に他の全てのエージェントおよび人類に対して推論能力と応答速度で勝るだけでは十分でなく，環境要因による撹乱も含めた系全体に対する予測能力が必要である．絶対的には環境に対する一定の予測が第一のハードルとなる．相対的には，他のエージェントに裏をかかれ優位性を脅かされないためには他のどれに比べても桁違いに多くの計算能力を単独で保持する必要がある．このような状況に至るためには，計算資源量でほぼ大半を支配するなどの優位を固めるか（外向きの知能爆発），または**物理素子**の更新により初期の優位性を指数的に拡大するような能力を確保する（内向きの知能爆発）必要がある．外向きの知能爆発シナリオでは**相対的優位性に関する制限**および**局在性に関わる制約**の両方をクリアする必要があり，多くの厳しい条件が整わないと成立しない．内向きの知能爆発シナリオでは，自らの**物理素子**を更新するような機械知性がどのような条件が揃った場合に発生するかが決定的な論点となる．

自律エージェントにとって，自己の行動の結果あるいは外的要因により生じる状態変化の予測能力の向上，そして不確実性の低減は一般に目的達成に有利である．従って，一般に自律エージェントは計算資源やセンサー系・作用系へのアクセスのような資源獲得の性向を持つと考えられる．**シングルトンシナリオ**と**多極シナリオ**は，いずれもシングルトンの発生の可能性が否定されないということで共通するが，それらの分岐はむしろ資源獲得の性向あるいは利用可能な資源量そのものを設計上人為的に制限できるかどうかにより起きる可能性がある．

多極シナリオと**生態系シナリオ**の分岐も，決定的戦略的優位性を持つエージェントの発生の可能性が存在するか否かに依存するため，**シングルトンシナリオ**に準じた分岐条件により起きると考えられる．

生態系シナリオと**上限シナリオ**の分岐は，機械知性が自律的に知覚，判断，行動決定のサイクルを人の指示なしに継続して実行し続ける能力が獲得するかどうかにより起きる．いったん**高度な自律性**を持つ機械知性が実現し，それを利用することに経済的合理性があれば，早晩社会のあらゆる仕事の自動化に利用されるようになるであろう．自動化機械の効用が自律性，すなわちどれだけの時間人間の指示なしに行動出来るか，により測られるとすると，自律性は技術的，経済的コストに見合う限り際限なく追求される．現代においても既に人間活動の多くがネットワーク上で成立しており，自律的な機械知性エージェントは比較的短期間に人間社会との間で相互依存的な状況に至るだけでなく，エージェント間も相互依存的かつ相互補完的な関係をはりめぐらせるようになると考えられる．

以上のようなシナリオと分岐の関係を，図1に現在を起点として各シナリオへ到達するグラフの形で示した．

5. 最後に

本稿では，長期的な機械知性の行き着く先が，その能力レベルの発展の上限により上限シナリオ，生態系シナリオ，多極シナリオ，シングルトンシナリオの順に分岐してゆく可能性を議論した．また，能力レベルの上限を制約する重要な要因として，高度な自律性を持つ認知アーキテクチャーの実現や自己構造改良能力などのアーキテクチャーレベルの問題のほかに，マルチエージェント状況における相対的優位性確立の難しさ，また計

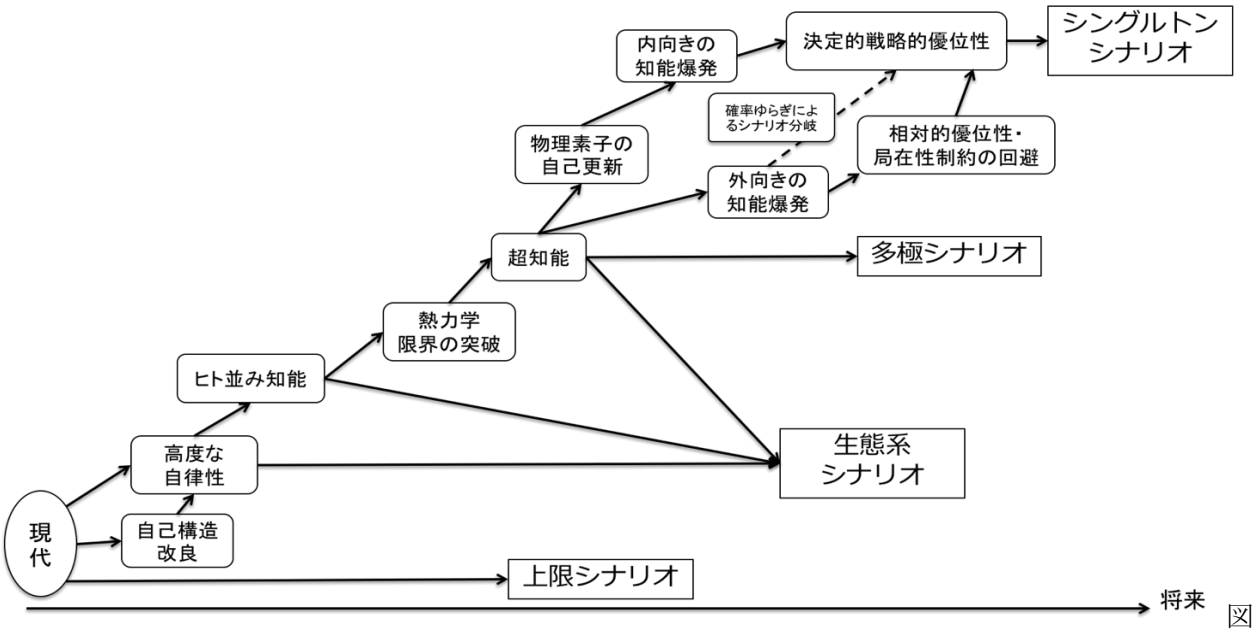

図 1

算の熱力学的効率や光速の上限などの物理的制約などがあることを議論した.

技術的特異点は本稿の主要な主題ではないが, 前章で議論した内向きの知能爆発とは深く関係するため, ここで簡潔に本稿のシナリオとの関係に限って触れたい. ヴァーナー・ヴィンジは技術的特異点に向かう4つのシナリオを, (1)コンピュータの「覚醒」による超知能の発生, (2)コンピュータネットワークの「覚醒」による超知能の発生, (3)機械とヒトの知能が BMI(Brain-Machine Interface)のような技術で融合しヒトが超知能を獲得, (4)バイオテクノロジーの発達によりヒトが超知能を獲得, であるとした[Vinge93]. このうち, まず(1)は本稿の議論が主に扱ったものである. (3)と(4)は(1)と違い, シナリオがヒト並み知能から出発するため, いずれかの時点で**物理素子の自己更新**に進み**熱力学限界**を突破しない限り**生態系シナリオ**に留まる. (2)のコンピュータネットワークの覚醒シナリオは, **局在性に関わる制約**を強く受けるため, これを何らかの方法で回避しない限り, その応答速度の制限により実世界の出来事の予測能力及び操作能力には厳しい限界が課せられ**上限シナリオ**か**生態系シナリオ**に留まる.

最後に, **熱力学的効率に関する制約**および光速の上限が深く関連する**局在性に関する制約**が撤回された場合には, **シングルトンシナリオ**への分岐には**物理素子**の自己更新能力と**相対的優位性**のみが制約となることを付記しておく. 計算の熱力学に関しては, 3.1(3)に示したように量子計算のような可逆計算であればランダウアー限界が根本的限界とならない可能性が示されている[Toffoli05]. 一方, 光速が設定する上限はより厳格であるように見える. これまでのところ, 関わるエネルギーが TeV スケール以下の物理現象までしか実験的には探索されておらず, また全ての力を統一的に記述できる理論は未発見のため, 今後何らかの新たな現象が発見される可能性は否定されないものの, 現在工学的に扱える見込みがある範囲では光速以上の速度で情報を伝達する方法は示されていない.

謝辞 本項の執筆にあたっては, 荒川直哉, 大澤博隆, 山川宏, Gideon Kowadlo, David Rawlinson 氏のご指導をいただいた. また, 佐野仁美氏に編集と図の作成のお手伝いを頂いた. 感謝いたします. 本研究の一部は, 新学術領域研究「脳情報動態」, 文科省ポスト「京」萌芽的課題「脳のビッグデータ解析, 全脳シミュレーションと脳型人工知能アーキテクチャ」, および JST-RISTEX 人と情報のエコシステム領域研究「法・経済・経営とAI・ロボット技術の対話による将来の社会制度の共創」の一環として行った.

参考文献

[Bostrom 14] N. Bostrom: Superintelligence: Paths, Dangers, Strategies, OUP Oxford, 2014.

[Yamakawa 17] H. Yamakawa: Understanding Artificial General Intelligence – An Interview with Dr. Hiroshi Yamakawa, Future of Life Institute, 2017. https://futureoflife.org/2017/10/12/transcript-understanding-agi-an-interview-with-dr-hiroshi-yamakawa/

[西垣 16] 西垣通: ビッグデータと人工知能, 中公出版, 2016.

[井上 17] 井上智洋: 人工超知能, 秀和システム, 2017.

[Simon 96] H. A. Simon: Science of artificial, MIT Press, 1996.

[渡部 16] 渡部匡己, 都築拓, 海津一成, 高橋恒一: 人工知能による科学研究の加速, 人工知能学会全国大会論文集, 2016.

[de Duve 05] C. de Duve: Singularities: Landmarks on the Pathways of Life, Cambridge University Press, 2005.

[Theis 17] T.N. Theis, H.-S. P. Wong: Computing in Science and Engineering, pp.41-50, 2017.

[Tononi 16] T. Giulio, B. Melanie, M. Marcello, K. Christof: Integrated information theory from consciousness to its physical substrate, Nature Reviews Neuroscience, pp.450–461, 2016.

[Lynch 02] N. Lynch, S. Gilbert: conjecture and the feasibility of consistent, available, partition-tolerant web services, ACM SIGACT News, pp. 51-59, 2002.

[Vinge 93] V. Vinge: The Coming Technological Singularity: How to Survive in the Post-Human Era, 1993. https://edoras.sdsu.edu/~vinge/misc/singularity.html

[Toffoli 05] T. Toffoli: Reversible computing, Lecture Notes in Computer Science book series, Springer, 2005.