特　集「人工知能と Emotion」

試論：人はなぜ感情をもつのか
―行動決定における感情の計算論的役割―

An Essay: Why We Have Emotions
― Computational Role of Emotions in Action Decision ―

大森　隆司　玉川大学脳科学研究所
Takashi Omori　Brain Science Institute, Tamagawa Univeristy.
omori@lab.tamagawa.ac.jp

Keywords: feeling, emotion, action decision, value calculation, computational theory.

1. はじめに：高等動物に共通の感情

感情は高等動物にはかなり普遍的な内部過程である．基本的に爬虫類以降のすべての動物は何らかの感情あるいは情動[*1]をもっている．その特徴は進化の過程で共通な部分が多いことで，多くの動物の感情表現は人間の考え方で解釈可能である．おそらくは共通の原理に基づいて動いているものと推測される．では，我々はなぜ感情をもっているのだろうか，なぜ感情は進化の過程で共通なのか，積極的な理由があるのではないか，というのが本稿の基本的な問いである．

本稿ではそれを議論するため，以下のように議論を進めていく．まず次章では，感情とはどのようなものか，現状の感情に関わる現象の心理的・モデル的理解および生理的理解について触れる．次いでこの2章では感情というものの範囲に関するカルテットセオリーについて紹介する．続いて3章では，本稿の主題である感情の計算論的意義の試論について述べる．まず，その基盤となる意思決定について現状の認知科学・経済学的な価値計算理論について述べ，それが感情の現象とかなり近いことを指摘する．続いて本稿の本題となる「感情＝価値計算システム仮説」の妥当性について述べ，最後にそれを実現するに必要な計算システムの姿について論じる．そして4章ではこれまでの議論を踏まえ，仮説の脳の感情理論としての可能性，限界について論じてまとめる．

2. 感情とはどのようなものか

現時点での一般的な感情のモデルとしては，Ekman [Ekman 97] や Russell [Russell 80] の研究がある．前者は，表情筋の活動の分析からその典型パターンを導き出し，結果として表情の分類モデルを与えている．一般には怒り，嫌悪，恐れ，幸福感，悲しみ，驚きの6種類の基本感情があるとされている．それに対して Russel は感情の基本特性として「快–不快」と「覚醒–眠気」の二次元の空間を想定し，そこに人間の各種の情動・感情状態を配置して情動の特性を表している．いずれも現象の説明モデルであり，どういう内部過程がその感情をつくり出しているかを説明するメカニズムのモデルではない．そのため，恐怖のような生理的・遺伝子的に我々につくり込まれていると思われる情動と，驚きのように予測という知的能力があって初めて可能となるより高度な情動が同じレベルで記述されている．

一方で，心理学は人間の感情現象を詳しく分析し考察してきた [北村 06, 北村 08]．心理学は行動実験による客観的な事実から感情の効果・影響について考察し，多くのことを発見してきた．しかしその方法では感情は何のためにあるのか，という問いに答えるには先が遠いように感じられる．例えば，Bower のネットワーク活動伝搬モデルは感情の内部過程に近寄ってはいるが [Bower 81]，計算論的な役割は明らかではない．

一方で，agi 社の光吉俊二氏は人間の高次の感情のモデルをつくり [光吉]，ロボット Pepper に組み込んだとされている [NHK 16]．興味深いものであるがモデルの内部過程が非公開であり，我々としてはこれ以上の議論はできない．

人間のより高次の感情についてモデルは，戸田の研究がある [戸田 92]．戸田は感情の目的は行動決定のための価値の計算にあるとしており，多くの複雑な感情，例えば嫉妬や愛情などを生存や生殖に関する基本的な価値に帰属させる推論過程を重視した．また，感情の非合理的な側面は，感情が進化した文明以前の野生時代には適応的であった情動に駆動される行動戦略が，文明の発達

*1 情動は感情の基本的な部分であると考えられるが，本稿では主に人間以外の感情を扱うこととして，感情と情動は区別しない．

により物理的な危険や時間圧の高い危険が少なくなったため不適応になったとしている．その内容は極めて魅力的であるが，やや強引な感もある．より精緻化したモデルへの展開が必要と思われるが，その基本原理は定性的な説明としては妥当性があるように思える．

ここまでは感情に関わる行動ベースの議論であった．次は情動のメカニズムとしての脳過程について考える．

脳における感情のシステムの研究は，ジョセフ・ルドゥーによるげっ歯類の恐怖条件付けの脳回路を研究によるところが大きい [ルドゥー 03]．これ以降，扁桃核は負の価値を認識する脳部位とされたが，最近では研究手法の進歩により正と負の両方の価値を表現するとされており，その構造についても新しい説が出ている [Janak 15]．

しかし，情動に関わる脳部位は扁桃核だけではない．例えば，我々の遺伝子に組み込まれているとしか思えない，ヘビに対する先天的な恐怖などがある．これに対するモデルとして Koelsch らにより感情のカルテットセオリーが提案されている [Koelsch 15]．このモデルでは，感情の対象を拡大して，身体の維持，安全の実現，愛着，経済的価値の 4 種類として，そのそれぞれに脳に対応する部位（脳幹，間脳，海馬，前頭眼窩野）があるとしている．

身体の維持とは，例えば食後の満足や空腹，腹痛などの基本的な身体状況とされている．通常，我々はこれを感情とは呼ばないが，その一部といっても間違いではないだろう．安全とは，恐怖や闘争心といった我々がいう感情に最も近いものである．愛着とは恐らくは記憶と感情が結び付いたもので，道具・環境・夫婦・共同体など多くの事例があるし，気に入ったものは多少のコストの犠牲を払っても獲得，維持したいと考える現象である．経済的価値を感情に入れることについては議論があろうが，例えばお金を得たときは嬉しい，損をしたらがっかりするなど，経済的価値と感情の結び付きは大きい．

一方で，我々の脳科学の知識からはカルテットセオリーが指摘した脳部位は狭いという気がする．例えば，正負の価値を符号化するとされる扁桃核，強化学習により行動の価値を学習するとされる大脳基底核，社会的行動に関わるとされる前頭葉内側部など，より多くの部位が感情に関わっているはずである．

3. 感情の計算的意義についての試論

3·1 価値計算に基づく意思決定

ここまで述べた感情・情動はどういう計算論的意味をもっているのだろうか．知能の価値は的確な行動決定にある．例えば，原始世界で住むヒトの祖先は，環境中の事物を認識して記憶し，それを後での行動計画などの的確な行動決定に利用したであろう．その際，より高い認識能力，記憶能力，推論能力が生存に有利であったことは疑いようがない．それが知能を生み出した必要性であったであろう．

ここで，意思決定について考える．身体がある知能の場合，意思決定とは行動決定に等しい．人間の意思決定に関しては主に認知科学や経済学で研究されており，そこでの基本概念は価値の計算である．意思決定とは，可能な選択肢のうち一つを選ぶことであり，経済学では我々はそれぞれの選択により得られる報酬，すなわち選択肢の価値を計算し，より価値の高い行動を選ぶとされている．この理論はもとは経済行動が対象であり，そこでは人間は一つ一つの価値を正しく計算できる合理的な存在とされている．認知科学においてもこの概念は同様で，人間は経済行動より幅広い行動において経済的価値とはまた異なる価値を，意識的，無意識的に計算・比較して意思決定するとされている [スタノヴィッチ 08]．

一方で，人間は合理的とは限らないことを示したのが行動経済学である．例えば，選択肢の成功確率が偏った場合や報酬が得られる時期が遠い未来の場合，人の価値の計算にはバイアスがかかることが知られており，その意味で人間は非合理的な存在である．それでも，人間は価値の予測に従って行動決定をする，という意味では上記の意思決定に関する基本概念は正しいものと考える [モッテルリーニ 08]．

意思決定の価値計算的説明は人工知能でも同様であるのは周知のとおりである．例えば将棋は，可能な選択肢を何手か繰り返した先の状態を予測し，その状態のゲームとしての良さ，すなわち価値を評価して，最も価値が高いと予想される状態をつくり出す現在の行動を選択するという，価値の計算に基づいている．これは将棋に限らず，評価関数を最大化する行動を探索するという，現在の人工知能の基本概念である．同様のことは学習について言え，誤差の最小化を目標として学習するニューラルネットワークは誤差最小という価値を実現するアルゴリズムを使用するし，強化学習は将来に期待される報酬の最大化を目的に学習する．すなわち価値の計算は，経済的場面に限らず評価関数の最大化という意味でより広い場面に適用可能な知能の原理である．

ここで改めて，カルテットセオリーの提唱する，身体·安全・愛着・経済的価値というより幅広い対象への感情の範囲の拡大と価値の関係を見てみよう．

[**身体**] 自身の身体維持は，そもそもの生物の本質であり，その実現に高い価値が割り振られるのは当然である．我々の経済活動はそもそも食べて身体を維持するためであり，我々はこの目的のために日々にお金，すなわち価値を支払っている．

[**安全**] 自身や家族・共同体の安全もまた種の保存の価値そのものである．安全を失うことは大きな損失につながり，我々は安全を他の利益を失ってでも獲得したい価値と考えている．

[**愛着**] 長く使ってきたような愛着をもったモノを我々

は捨てがたい．愛着には明らかに価値に相当するものが付随している．愛着のメカニズムは知られていないが，例えばあるものを使った記憶に好ましさが連合され，そのものの想起や認識に付随して自動的に価値が想起される，という説明があり得よう．

［経済的価値］お金は他の具体的な価値を手に入れるための媒体である．その媒体そのものには価値はないが，我々の認識にはお金は価値そのものとして浸みついている．

すなわち，カルテットセオリーが感情の要素と考える身体，安全，愛着，経済的価値はすべて価値に関わっており，経済学が対象として前頭葉が計算する金銭的価値と合わせて，感情の過程は広義の価値として検討可能である，ということができよう．そして価値の計算とは意思決定の基本原理である．すなわち，広義の感情とは価値の計算システムそのものではないか，という仮説が生まれる．

3·2 感情＝価値計算システム 仮説の妥当性

仮説「広義の感情とは，意思決定のための価値計算システムである」

この仮説によると，カルテットセオリーが述べた広義の感情の4成分（身体，安全，愛着，経済的利益）に対応する種類の異なる価値が存在するはずである．身体についての価値は我々の生存に直結する状態であり，価値の評価システムとは我々の体，すなわち遺伝子に組み込まれていると考えるべきであろう．例えば，寒い，空腹，悪寒，といった自己身体の状態の認識はそれだけでは感情とはいわないが，明らかに我々のムードに影響するし，それを解消するために他の価値を犠牲にすることも多い．端的に言うと，我々は空腹になると機嫌が悪くなる．これからも，身体状態は広義の感情に含まるといえよう．

安全に関わる感情については，ヘビに対する恐怖のように進化の過程で遺伝子に組み込まれて脳でprewiredな状態であると考えられる感情がある．また，喜怒哀楽や闘争・恐怖のような典型的な感情の多くは世界中で文化によらず共通であり，遺伝的に決まったprewiredな部分が大きいと考えるべきであろう．このうち少なくとも恐怖や闘争といった感情は，個人・家族・共同体の安全を高める価値があることは明らかである．身体維持に関わる感情との違いは，感情の原因が外界にあることで，結果的に社会的な価値も含むであろう．

安全に関わる感情では，それに学習的に獲得したものが追加され，その学習の責任部位が扁桃核であるとされている．一般に感情を符号化する脳部位は扁桃核であるとされるが，ここで議論している広義の感情では扁桃核はその一部，安全に関わる感情の学習的獲得に関わると考える．学習が必要なのは，感情の原因が外界にあるため必然で，変化する環境に適応するためのシステム的戦略であろう．

愛着とは一般には事物・人物に対するポジティブな感情とされている．しかしそのような特定の対象がポジティブな感情に結び付くには，それなりの経験が必要であろう．すなわち，愛着には記憶と感情の連合が必要であると考える．しかもそれが単一のエピソードを超えて長期間の繰返しによって習慣化（Habit）されて定着したとき，愛着と呼ばれるものになると考える．そのようなHabitにより，人がその事物・人物に接したときに無意識のうちにそれに対する感情が付随して想起される，という現象をここでは愛着と考える．おそらく嫌悪も同様の現象であろう．愛着·嫌悪の利点は，その個体にとってポジティブ・ネガティブな行動を意識せずとも自然に取る・回避するようになるということで，生存確率の上昇や安全確保に寄与することは間違いない．カルテットセオリーではその責任部位を海馬としているが，海馬以外にも同様の機能を担う脳部位はありそうで，その脳での実態の解明はこれからであろう．一方で人工知能でこれに対応する方法としては，例えば強化学習はこのような機能をもっているし，他の報酬最大化を行う学習はいずれもこの機能を果たし得る．そして何より，これまでなかなか人工知能的には説明の難しい「愛情」についても議論の可能性が出てくる，というのが愛着の注目するべき点であろう．夫婦関係や恋については戸田がその価値を検討しているが，洗練の余地があるように感じられる [戸田 92].

経済的な利益の計算が感情に直結するというのは，例えば株で損して怒る，儲けて喜ぶ，など枚挙にいとまがない．そもそもお金とは価値を抽象化したものであり，お金により我々の生存可能性や身体の安全性，さらに愛着を通じた価値の追求も可能となる．その意味で，経済的利益が広義の感情に含まれるのは当然である．カルテットセオリーはこの責任脳部位を前頭葉眼窩野としている．これは臨床例などからも間違いではないが，実際の我々の経済的利益の計算は前頭葉のもっと広い部位で行われているように思える．

例えば「情けは人のためならず」というように，我々は自身の価値を多少犠牲にしても他者を助けることが結果的に自分の価値につながると考える．この背景には社会という複雑な体系があるのであるが，そこに関わる複雑な関係を推論して現在の状況に対する価値，さらには将来的な価値を予測することは，人間の社会性の重要な機能であろう．その責任部位は，前頭葉の内側面とされており，その範囲はかなり広いと推定される．そしてその一例として，他者の感情状態を無意識のうちに推定してしまうエンパシーであり，さらにより知的な機能が加わってシンパシーがあると考えると，その存在意義が理解しやすい．

以上，広義の感情を価値計算システムと考えることの妥当性について述べてきた．進化の過程で少なくとも哺乳類，おそらくは爬虫類以降，生存に関わる情報とその

ための処理システムは重要な価値をもっており，幅広い種での類似の表現型としての感情表現が進化してきたのではないか，というのがその起源についてのイメージである．

しかし例外はヒトである．ヒトの感情は生存に必要という以上に複雑化しているように思える．ヒトの場合，戸田が指摘したように感情が進化した非文明の野生環境と，現在我々が暮らしている文明環境の乖離がまずある．それと同時に，ヒトには本来の単純な価値評価システムに加え，高度に進化した推論システムと社会化に対応した状況認識機能が組み込まれている．それにより，例えば軽蔑，称賛，尊敬，感謝，罪悪感，恥といった我々自身にも説明が難しい感情の内部処理メカニズムが加わったのであろう．さらにこの延長に笑いや遊びといった謎がある．現時点では価値計算システムという視点から人間の豊かな感情現象を説明するアイディアは限られている．

3·3 感情＝価値計算システムの計算論的モデル

感情を価値計算システムであるとしたとき，それはどういう計算過程で実現されているのだろうか．感情の計算メカニズムの理解は人工知能としてその実現可能性に関わり，重要である．表 1 にそのあり得る計算モデルについてまとめた．身体と安全についてはかなりの部分が固定であり，学習が含まれる部分も比較的容易なパターン認識で実現可能と思われる．一番の課題は外界の事物·事象・状況の認識であろう．

表 1 価値計算システムのあり得る計算モデル

項　目	あり得る計算モデル	該当する脳部位
身　体	センサによる直接検出およびその組合せによるパターン認識と固定価値	遺伝子で固定された生体センサによる検出とパターン認識
安　全	感覚センサからの固定パターンおよび学習されたパターンの認識と価値判断	感覚領野，特徴抽出系としての感覚性連合野，間脳，扁桃核
愛　着	場面認識と価値の連合（エピソード）と，その一般化（強化学習）	新皮質での場面認識，海馬系での記憶，線条体
経済的価値	知覚された現在状況からの推論，ツリー探索や関数近似での状況–価値マッピング	前頭眼窩野，社会的価値については前頭葉内側面，推論は前頭葉とされる

愛着については，個々の場面での価値につながる事物の認識と，その場面の集合の一般化による Habit の形成という 2 段階があろう．前者はパターン認識で，後者は強化学習で説明できる．こう考えると，愛着については基本的にはこれまでの機械学習の枠組みでのアプローチが可能であるように見える．ここまでの計算理論はすでに多くの研究があり，そのどれを使うか，あるいは場面や対象に応じたアルゴリズムの選択が課題であろう．

残る，価値の推論と，全体システムの統合制御については，議論が必要である．経済的な場面に限らず広い意味の価値の推論とは，以下のようにモデル化できよう．

(1) 新奇場面に出合って強化学習などの過去の経験による価値付けができない場合，
(2) 環境についての知識に基づき起き得る場面を予測することを反復し，
(3) 予測の結果に価値付けが可能な状況に到達し，
(4) 予測の不確実性を考慮して (1) の新奇場面についての価値を割り当てる．
(5) 予測可能なすべての状況について価値を求め，価値が最大化する現在の行動を選ぶ．

この方法は，人工知能の分野ではおなじみの Tree 探索による予測と評価の過程そのものである．例えば将棋では，現在の盤面で可能な手を一つ選んで何手か先までの自己と他者の手を価値最大化の戦略により予測し，結果として得られた盤面の強さを評価する．上記の説明は，それと同じことをより一般的な場面について行う，ということである．従来，人工知能ではこの推論を論理的な計算過程としてモデル化してきた．しかし現在，人間の推論が本当に論理なのかは明らかではないし，脳におけるシンボル的な処理過程も未解明である．価値の推論を論理推論として表現することについては，再検討が必要であろう．

4. まとめと議論

本稿では，感情とは脳の価値計算システムであるとする仮説について論じてきた．関連する分野として，人間の意思決定の理論，脳の感情システムの生理学，4 種類の価値の感情としての妥当性，人工知能としての実装のためのあり得る計算理論について論じた．感情と価値が結び付くことは直観的には理解していただけると思うが，その価値計算システムとしての生理学的，計算論的検証は不十分で，それを支える事実を集める努力が必要である．

この仮説の適用範囲は，ヒトを除く哺乳類と，恐らくは爬虫類も含むであろう．問題はヒトで，現時点では本仮説はヒトの感情はその高度の知的情報処理と組み合わさっているため複雑で説明しきれない，という立場である．しかも，戸田が言うように，感情は非文明化時代に進化・適応した機能であり，現代の文明化社会に対する適応度は高いとは言い切れない．モデル化にあたり，さらには人工知能としての実装にあたり，このずれをどうするべきであろうか．これらについては多様な立場からの検討が必要である．

本仮説によると，感情とは脳全体の意思決定のシステムの表現型である．知能とは価値の計算を高度化して個体および種の生存をより確実にする道具である．現在の人工知能はその根源としての価値の議論をしないまま，

表現型の計算過程のモデル化に力を入れているように見える．ひょっとすると知能の現象の本質を見誤っているのではないか，という気もする．逆に，感情のみを取り出してモデル化することもまた適当ではない，というのも本仮説が示唆することである．

この議論の先に，何に価値を置くかという価値観の議論があろう．価値観は，個人，組織，社会，文化のそれぞれにおいて共有される事物の優先順位，重み付けの体系とされている．我々が感情と呼ぶものがヒトの社会的環境の中での価値の表現手段として進化したものであるなら，極めて社会的な動物であるヒトの感情が他の動物と比較して特に複雑であるというのもうなずける話ではある．そしてその価値観を言語や論理で体系づけようとしたものが，倫理やモラルであろうか．ここは専門家の意見をうかがい，本仮説の位置付けを深めたい．

人類と人工知能の関係についても同様に議論が必要である．人工知能は人類を滅ぼすかという議論は生産的ではない．人工知能の価値観として何を埋め込むか，そのための価値の計算方法はどういうものが可能かを論じることが，人間が安心して付き合える人工知能の設計論として必要と思う．ひるがえって現在の多くの人工知能研究では，学習の評価関数は人間が与えている．ところが，価値や倫理についての明示的な検討をしないままの報酬や評価関数の設計は，設計者が意図せずして人類にとって危険な価値を人工知能に埋め込む可能性は排除されない [Hibbard 15]．自己組織化という方法も同様である．自己組織化とはシステムが行動を変えるための学習式の中に暗黙裏に価値が埋め込まれた学習方式であり，結果がどこに向かうかは明示的に示されないだけにより判りにくい．注意が必要である．本試論が，人工知能の在り方についての一つの示唆になれば幸いである．

本研究は文部科学省科研費 15H01622 の助成を受けた．支援に感謝する．

◇ 参 考 文 献 ◇

[Bower 81] Bower, G. H.: Mood and memory, *American Psychologist,* Vol. 36, pp. 129-148（1981）
[Ekman 97] Ekman, P., et al.: What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system（FACS）, *Series in Affective Science*（1997）
[Hibbard15] Hibbard, B.: *Ethical Artificial Intelligence,* arxiv.org/abs/1411.1373（2015）
[Janak 15] Janak, P. and Tye, K.: From circuits to behavior in the amygdala, *Nature Review,* Vol. 517, pp.284-292（2015）
[北村 06] 北村英哉，木村 晴 編：感情研究の新展開，ナカニシヤ出版（2006）
[北村 08] 北村英哉：感情研究の最新理論―社会的認知の観点から―，感情心理学研究，Vol. 16, No. 2, pp. 156-166（2008）
[Koelsch 15] Koelsch, S., et al.: The quartet theory of human emotions: An integrative and neurofunctional model, *Phys. of Life Reviews,* Vol. 13, pp. 1-27（2015）
[スタノヴィッチ 08] キース・E・スタノヴィッチ：心は遺伝子の論理で決まるのか―二重過程モデルでみるヒトの合理性，みすず書房（2008）
[モッテルリーニ 08] マッテオ・モッテルリーニ：経済は感情で動く―はじめての行動経済学，紀伊國屋書店（2008）
[NHK 06] NHK スペシャル：天使か悪魔か 羽生善治 人工知能を語る，http://www.nhk.or.jp/special/ai/（2016年5月15日放送）
[ルドゥー 03] ジョセフ・ルドゥー：エモーショナル・ブレイン，東京大学出版会（2003）
[Russell 80] Russell, J.: A circumplex model of affect, *J. Personality and Social Psychology,* Vol. 39, pp. 1161-1178（1980）
[戸田 92] 戸田正直：感情，東京大学出版会（1992）
[光吉] http://www.agi-web.co.jp/technology/index.html

2016年7月25日 受理

著 者 紹 介

大森 隆司（正会員）

1980年東京大学大学院工学研究科計数工学専攻修了．1981年より東京大学助手，1987年東京農工大学工学部講師を経て，1988年より同助教授，1998年より同大学電気電子工学科教授，2000年5月より北海道大学大学院工学研究科教授を経て，2006年4月より玉川大学教授，現在に至る．工学博士．この間，1989～90年ブラウン大学言語と認知学科客員研究員．脳という神経機構に知的な行動が生まれる情報的なメカニズムに興味があり，認知科学，人工知能，発達，神経科学などの諸学問を足を突っ込みながら，心に関わる脳の情報処理過程の解明と工学的な方法による実現を試みている．