論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	意思決定における確率の取り扱いに関する理論的及び実証的研究
著者(和文)	松森嘉織好
Author(English)	Kaosu Matsumori
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第11221号, 授与年月日:2019年6月30日, 学位の種別:課程博士, 審査員:小池 康晴,伊東 利哉,中村 健太郎,金子 寛彦,吉村 奈津江
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11221号, Conferred date:2019/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

東京工業大学　令和元年度　博士論文

意思決定における確率の取り扱い

に関する理論的及び実証的研究

東京工業大学大学院総合理工学研究科

物理情報システム専攻

指導教員：小池 康晴

松森 嘉織好

目次

第 1 章 序論

1. はじめに

この文章を読み続けるかどうかといった日常的な選択から、どのような仕事につくかといった人生の岐路における選択に至るまで、我々の生活は意思決定の連続である。望ましい意思決定とは何か、また実際に人やその他の動物がどのように意思決定しているかについて、工学、統計学、神経科学、経済学、経営学、心理学などさまざまな分野で研究されている (Gilboa 2010; Gilboa 2012)。

合理的意思決定の理論で最もポピュラーなアプローチは主観的価値である効用 U を最大化するように行動 a を決定するという効用最大化アプローチである。例えば、ある人が外界に関する「午後から雨が降る」、「傘をさすと雨に濡れない」の 2 つの信念と、「雨に濡れないことの効用が 10 」、「雨に濡れてしまうことの効用が 0 」という内的な効用を持っていたとする。外界に関する信念と内的な欲求を組み合わせて、効用を最大化するように行動を意図すると (Bratman 1987)、この場合、傘を持って出かけることになる。

2. 期待効用理論とベイズ推定

では、その日に雨が降るかどうかがはっきりとは分からない場合に、傘を持っていくかどうかを決めるにはどうしたら良いだろうか。どのような状態が実現するかが、

はっきりとは分からず、行動の結果が確率的に決まる環境における意思決定の基準としては、期待効用最大化の基準が最もよく採用されている (von Neumann and Morgenstern 1947)。これは、確率 P と主観的価値 U との積（期待効用 EU = P × U）が最大になるような行動 a を選択するというものである。確率 P の学習に関しては、ベイズ推定が最もよく採用されている。これは、外界から得られた新たな情報によって、これまでの確率 P に関する信念をどのように更新するかを記述するものである。期待効用理論と確率 P に関するベイズ推定とを組み合わせることで、さまざまな分野で別々におこなわれてきた意思決定研究は統合的に再解釈することができるため、近年では一つの研究分野と見なされるようになってきている。本論文では、通常の期待効用理論では説明できなかった現象を拡張された期待効用理論によって説明する試みをおこなう。

3. 本論文の構成

第 2 章では、期待効用理論に基づく意思決定モデルに、期待効用理論とは別の流れの中で発展してきた社会心理学の概念を組み込んだ。人々の望ましくない習慣や中毒的な行動を変える「行動変容」の問題は、さまざまな分野で長い間研究されてきた。多くの行動変容のモデルは、行動の意図を生じさせるための重要な要因として、態度、規範、自己効力感の 3 つを挙げるという点で一致している。一方、行動変容モデルの

精度を向上させるために、既存の行動変容モデルと行動経済学の成果を組み合わせる試みがおこなわれつつある。しかしながら、この試みは計画的行動理論などの既存の行動変容モデルが、多くの行動経済学的モデルの基盤となっている期待効用理論と整合的ではないため、あまり成功しているとは言い難い。第 2 章では、既存の行動変容モデルと期待効用理論の構成要素の対応関係を明らかにした上で、期待効用理論の自然な拡張としての決定理論的な行動変容モデルを提案する。なお、第 2 章の内容は、著者らによる *Frontiers in Psychology* に採択された論文の内容に基づいている (Matsumori, Iijima et al. 2019)。

第 3 章では、期待効用理論と確率 P に関するベイズ推定とを組み合わせた規範的モデルから、動物や人の意思決定が逸脱する場合について取り扱う。よく知られた逸脱として、人の確率推論にバイアスがあることが知られている。このバイアスを定量的に記述するために指数型バイアス付きベイズ推定モデルを考える。ベイズ推定の二つの構成要素である尤度と事前確率に関する指数バイアスをそれぞれ考え、二種類のバイアスの強さに基づく「バイアス平面」を考えることができることを示すとともに、バイアス平面上に最尤推定、ベイズ推定、maximum a posteriori (MAP) 推定を位置づけることができることを明らかにする。また、さまざまな人間の推論バイアスもバイアス平面上に位置づけられることを示す。さらに、既存のベイズ推定の神経モデルである確率的ポピュレーションコーディングモデル (Ma, Beck et al. 2006) におけるシナプス

入力のゲインを変更することが、指数型バイアスをかけることに対応していることを明らかにする。これにより、知覚的意思決定を説明する神経モデルをベイズ推定の立場から解釈する。最後に、認知制御が指数バイアスの強さを調整するメカニズムとして機能している可能性について考察する。なお、第 3 章の内容は、著者らによる *Frontiers in Neuroscience* に採択された論文の内容に基づいている (Matsumori, Koike et al. 2018)。

第 4 章では全体のまとめをおこなう。

第 2 章 社会心理学への適用

医療従事者やトレーナーにとって、人びとの食べすぎや過度の飲酒、運動不足、喫煙などの望ましくない習慣的、中毒的な行動を変えさせることは難しい場合も多い。どうすれば、我々は人々の行動を望ましいものに変えることを手助けすることができるのだろうか。このような「行動変容」の問題は、心理学や教育学、看護学、公衆衛生、医学、ヘルスプロモーションなどのさまざまな分野で長い間研究されてきた (Fishbein and Ajzen 2010)。多くの行動変容のモデルは、行動の意図を生じさせるための重要な要因として、態度、規範、自己効力感の 3 つを挙げるという点で一致している (Sheeran, Maki et al. 2016)。しかし、社会的認知理論や計画的行動理論といった既存の行動変容モデルは、行動の生成確率や介入による行動の生成確率の変化を十分な精度では予測できていない (Sniehotta, Presseau et al. 2014)。

行動変容モデルの精度を向上させるために、既存の行動変容モデルと行動経済学の成果を組み合わせる試みがおこなわれつつある (Roberto and Kawachi 2015)。行動経済学的なモデルはさまざまな人間の行動バイアスを分析対象としているため、行動経済学と行動変容の組み合わせは非常に有用だと考えられている。しかしながら、既存の行動変容モデルが、多くの行動経済学的モデル (Kahneman and Tversky 1979; Schoemaker 1982) の基盤となっている期待効用理論と整合的ではないために、二つの分野のモデルを組み合わせることはチャレンジングな試みとなっている。

本章では、主要な行動変容モデルである計画的行動理論と、期待効用理論の構成要素の対応関係を明らかにした上で、新しいモデルである決定理論的な行動変容モデルを提案する（図 2.1）。このモデルは、主観的規範と自己効力感のコンポーネントを通常の期待効用理論に付け加えたものになっている。

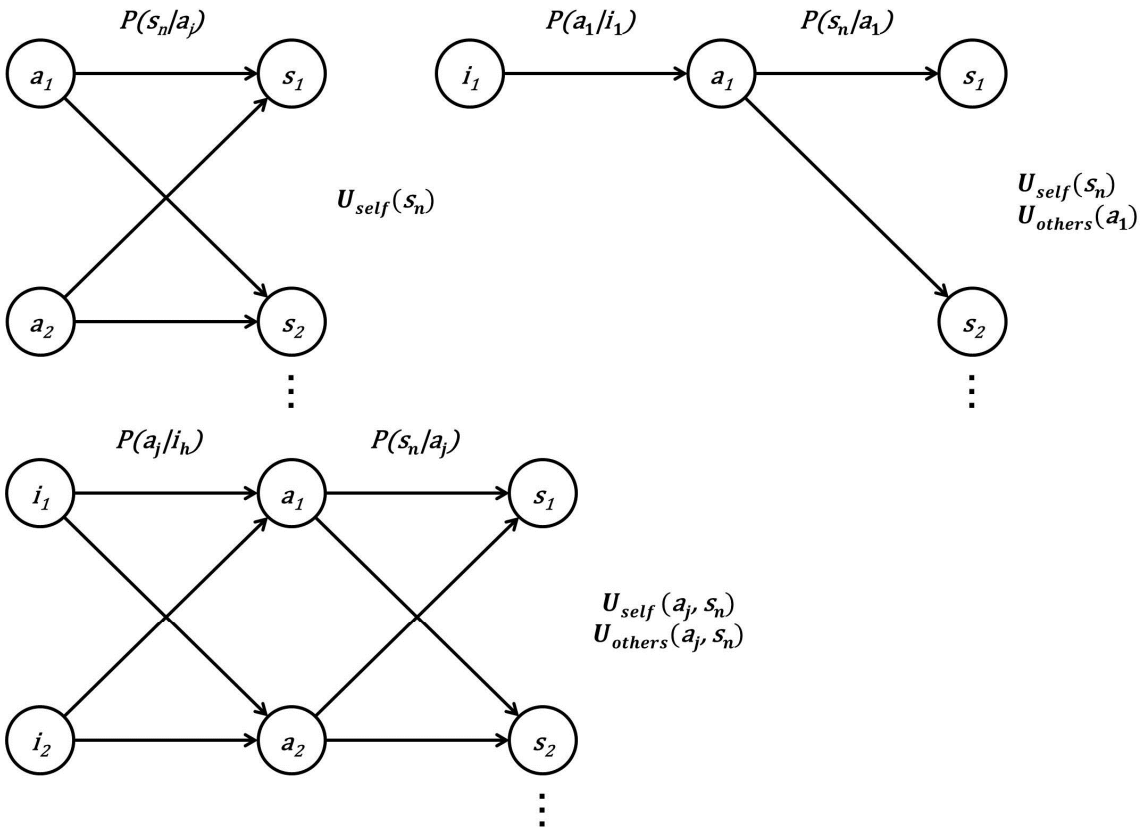

図 2.1 (A) 期待効用理論。期待効用理論は、確率的な環境下での合理的意思決定理論としては最もポピュラーなものである。状態集合 ($S = \{s_1, s_2, \ldots, s_n, \ldots, s_N\}$)、行動集合 ($A = \{a_1, a_2, \ldots, a_j, \ldots, a_J\}$)、行動 a_j をとった時に状態 s_n が生じる主観確率 ($P(s_n|a_j)$)、そして状態 s_n の主観的効用 ($U_{self}(s_n)$) が与えられた時、期待効用理論によると、エージェントは主観的効用の期待値 $E[U_{self}|a_j]$ を最大化するような行動 a_j を選択する。(B) 計画的行動理論の期待効用理論風の図式化。ターゲット行動をとろうという意図 (i_1) が追加的に導入されている。計画的行動理論では、行動への態度、主観的規範、知覚された自己効力感の 3 つの要因によって行動意図が決まる。行動への態度は、$P(s_n|a_1)$ と $U_{self}(s_n)$ によって決まる。主観的規範は $U_{others}(a_1)$、知覚された自己効力感は $P(a_1|i_1)$ である。(C) 決定理論的な行動変容モデル。このモデルでは意図集合 $I = \{i_1$: ターゲット行動をとろうと意図する, i_2: ターゲット行動をとろうと意図しない$\}$) を導入する。エージェントは $E[(U_{self} + wU_{others})|i_h]$ を最大化するように意図を選択する。

1 節では、期待効用理論の概要を行動変容の文脈に従って説明する。2 節では、計画的行動理論の説明をおこない、さらに計画的行動理論を決定理論的に解釈する。3 節では、新しいモデルである決定理論的な行動変容モデルを通常の期待効用理論を拡張する形で提案する。4 節では、提案モデルの優位性について議論する。最後に全体の内容を要約して、今後の方向性について議論する。

1. 通常の期待効用理論

期待効用理論は、確率的な環境下での合理的意思決定理論としては最もポピュラーなものである (von Neumann and Morgenstern 1947)。

状態集合 ($S = \{s_1, s_2, \ldots, s_n, \ldots, s_N\}$)、行動集合 ($A = \{a_1, a_2, \ldots, a_j, \ldots, a_J\}$)、行動 a_j をとった時に状態 s_n が生じる主観確率 ($P(s_n|a_j)$)、そして状態 s_n の主観的効用 ($U_{self}(s_n)$) が与えられた時、期待効用理論によると、エージェントは主観的効用の期待値を最大化するような行動 a_j を選択する。

$$\boldsymbol{E[U_{self}|a_j] = \sum_{n=1}^{N} P(s_n|a_j)U_{self}(s_n)}$$

(式 2.1)

本章では、行動集合が 2 つの相補的な要素からなる場合 ($A = \{a_1$: ターゲット行動を実行

する, a_2: ターゲット行動を実行しない}) を考える (図 2.1A)。

多くの経験的な研究で、エージェントの行動選択ルールとして、以下のようなシグモイド関数によるものが採用されている (Luce 1959; Sutton and Barto 1998)。

$$\boldsymbol{P(a_1) = sigmoid(\beta_1 \cdot \{E[U_{self}|a_1] - E[U_{self}|a_2]\} + \beta_0)}$$

(式 2.2)

ここで、逆温度 β_1 は行動選択のランダムさの度合いを調節し、定数 β_0 は決定のバイアスを表している。

例えば、状態集合 S = {s_1: 健康, s_2: 病気}、行動集合 A = {a_1: 運動する, a_2: 運動しない}、そしてエージェントの持つ信念及び効用として、$P(s_1|\ a_1) = 0.8$, $P(s_1|\ a_2) = 0.2$, $U_{self}(s_1) = 1$, $U_{self}(s_2) = 0$ というような状況を考えてみよう。このとき、それぞれの行動の期待効用は、

$$\boldsymbol{E[U_{self}|a_1] = \sum_{n=1}^{2} P(s_n|a_1)U_{self}(s_n) = 0.8 \cdot 1 + 0.2 \cdot 0 = 0.8}$$

$$\boldsymbol{E[U_{self}|a_2] = 0.2 \cdot 1 + 0.8 \cdot 0 = 0.2}$$

となる。さらに、エージェントの逆温度が $\beta_1 = 1$、かつ定数項が $\beta_0 = 0$、のとき、期待効用理論では、このエージェントが運動する確率を $P(a_1) \fallingdotseq 0.65$ と予測することになる。

2. 計画的行動理論

計画的行動理論は典型的な行動変容モデルであり、このモデルによると、ターゲット行動 (a_1) の行動意図は、行動への態度、主観的規範、自己効力感の 3 つの要因によって決まるとされる (図 2.2)。計画的行動理論を用いる際には、これらの要因は質問紙により測定される。

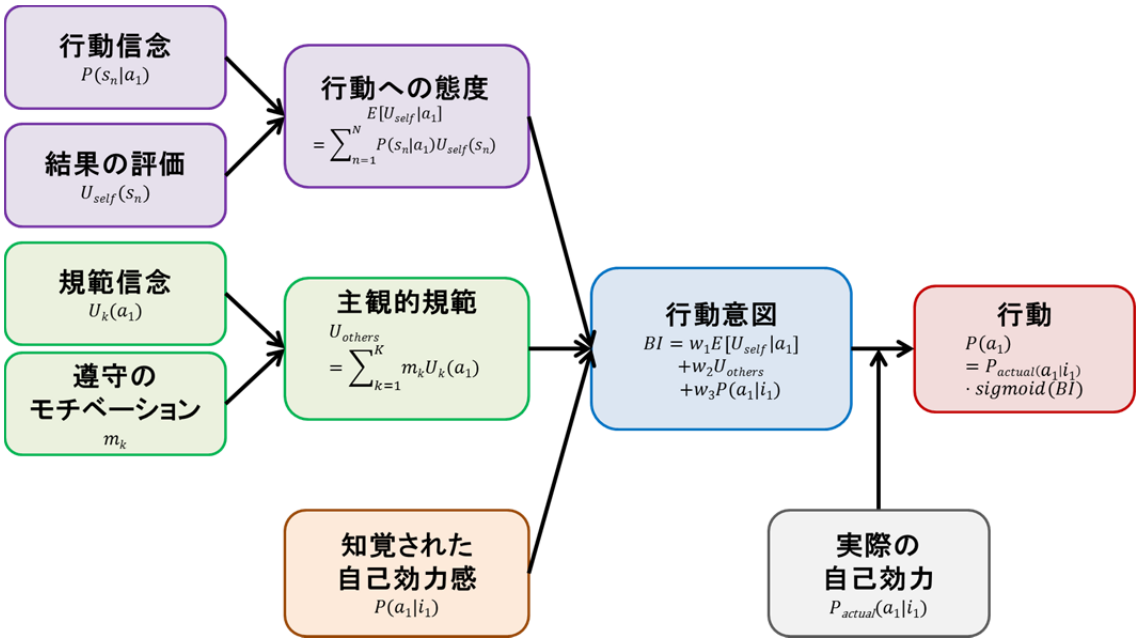

図 2.2 計画的行動理論。計画的行動理論は典型的な行動変容モデルであり、このモデルによると、ターゲット行動 (a_1) の行動意図は、行動への態度、主観的規範、自己効力感の 3 つの要因によって決まる。行動への態度 ($E[U_{self}|a_1]$) は、行動信念 ($P(s_n|a_1)$) と結果の評価 ($U_{self}(s_n)$) を掛けたものの和によって決まる（紫）。主観的規範 (U_{others}) は、規範信念 ($U_k(a_1)$) と遵守のモチベーション (m_k) を掛けたものの和によって決まる（緑）。知覚された自己効力感は、ターゲット行動をとろうとした際に実際にどのくらい上手く実行できるかに関する信念 ($P(a_1|i_1)$) を意味する（オレンジ）。行動意図（青）はこれら 3 つの重みつき和によって決まる。行動の生成（赤）は行動意図と実際の自己効力 ($P_{actual}(a_1|i_1)$) の関数である（グレー）。

行動への態度はターゲット行動 a_1 を実行することに対するエージェントの評価を意味する (Ajzen 1991; Fishbein and Ajzen 2010)。これは、経済学における期待効用理論、心理学における期待価値理論に基づいた概念である (Edwards 1954; Ajzen 1985)。

行動への態度は、「行動信念」と「結果の評価」を掛けたものの和によって決まる。行動信念とは、ターゲット行動を実行することがどのような結果 s_n を引き起こすかということに関するエージェントの信念（主観確率）のことである (Ajzen 1985)。そのため以下では、行動信念を $P(s_n|a_1)$ と表記することにする。 また結果の評価とは、ある結果がもたらされたときのエージェントの効用 $U_{self}(s_n)$ のことである (Ajzen 1985)。これらのことから、行動への態度はターゲット行動がとられた際の、期待効用のことであると考えることができる ($E[U_{self}|a_1] = \Sigma_{n=1}^{N} P(s_n|a_1)*U_{self}(s_n)$) (Edwards 1954; Ajzen 1985; Fishbein and Ajzen 2010)。ここで、期待効用理論では $E[U_{self}|a_1]$ と $E[U_{self}|a_2]$ の両方が考えられていたのに対し、計画的行動理論では $E[U_{self}|a_1]$ しか明示的に考えられていない。

行動への態度のみでは、エージェントの行動を十分説明できなかったことから、計画的行動理論には、他の 2 つの要因（主観的規範と自己効力感）が付け加わっている。

主観的規範は、行動をとるかどうかについての知覚された社会的プレッシャーのことである (Fishbein and Ajzen 2010)。主観的規範は、「規範信念」と「遵守のモチベーション」を掛けたものの和によって決まる。規範信念はある個人 k (k = 1, 2, …, K) がどの程度自分

にターゲット行動をとってほしいと考えているとエージェント自身が考えているかを意味しており、これは他者の効用 $U_k(a_1)$ の推定値とみなすことができる。エージェントが個人 k の効用をどの程度重要視するか決める係数を遵守へのモチベーションと呼ぶ (m_k; k = 1, 2, …, K)。まとめると、主観的規範は各個人の効用の重みつき和とみなすことができる ($U_{others}(a_1) = \Sigma_{k=1}^{K} m_k * U_k(a_1)$) (Fishbein and Ajzen 2010)。ここで、計画的行動理論では、他者の効用の定義域は行動であるのに対して、行動への態度におけるエージェント自身の効用の定義域は結果の状態であり、二つの効用関数の定義域は異なったものになってしまっている。

（知覚された）自己効力感は、ターゲット行動をとろうとした際に実際にどのくらい上手く実行できるかに関する信念を意味する (Bandura 1982)。 自己効力感の概念の提唱者であるバンデューラは行動と結果に関する信念である結果予期に加えて、自己効力感が人間の行動の決定要因として重要であると強調した (図 2.3)。 知覚された自己効力感は、エージェントがターゲット行動をとろうと意図した場合 (i_1) に、成功裏にターゲット行動を実行できる確率に関する信念のことであるため、$P(a_1|i_1)$ と書くことができる。ここで、結果予期は行動への態度の説明に出てきた行動信念と同一の概念である。

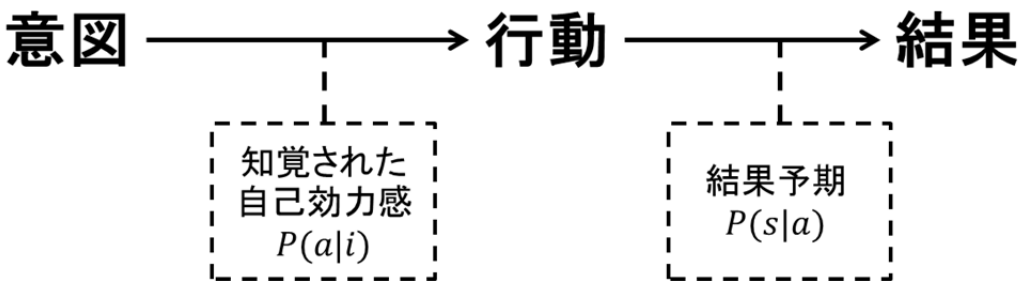

図 2.3 バンデューラの図式。結果予期だけではなく知覚された自己効力感が人間行動の決定要因であると考える。知覚された自己効力感は、行動をとろうとした際に実際にどのくらい上手く実行できるかに関する信念 (P(a|i)) である。結果予期は、ターゲット行動がどのような結果を生むかに関する信念 (P(s|a)) である。

3 つの概念（行動への態度、主観的規範、知覚された自己効力感）の重みつき和が行動意図 (Behavior Intention, BI) を決める (図 2.2)。

$$\boldsymbol{BI(i_1) = w_1 E[U_{self}|a_1] + w_2 U_{others}(a_1) + w_3 P(a_1|i_1)}$$

(式 2.3)

ここで、w_1, w_2, w_3 は行動への態度、主観的規範、知覚された自己効力感それぞれの重み係数である。この等式を次の節に出てくる提案モデルと比較するために書き換えると、

$$\boldsymbol{BI(i_1) = w_3 P(a_1|i_1) + \sum_{n=1}^{N} P(s_n|a_1)\{w_1 U_{self}(s_n) + w_2 U_{others}(a_1)\}}$$

(式 2.3')

のようになる。式 2.3 の第二項と U_{others} についての 式 2.3' の対応部分は、$\sum_{n=1}^{N} P(s_n|a_1) = 1$ であることから一致する。

ここで、主観的規範と自己効力感が $E[U_{self}|a_1]$ に単純に足されているため、計画的行動理論における行動意図は期待効用理論と整合的ではない。言い換えると、行動への態度に主観的規範と自己効力感の概念を加えることで精度を増そうとした計画的行動理論の試みは、行動経済学的モデルの根底にある期待効用理論との整合性を失わせるものになってしまっている (Kahneman and Tversky 1979; Schoemaker 1982)。理解を助けるために、期待効用理論の枠組みで計画的行動理論を図式化しようと試みた (図 2.1B)。

行動集合が二つの要素からなる時 ({a_1: ターゲット行動を実行する, a_2: ターゲット行動を実行しない})、エージェントの意図推定のためにロジスティック回帰がよく用いられている。これは、エージェントの意図選択ルールがシグモイド関数に基づいていると仮定することに対応している (Luce 1959; Sutton and Barto 1998)。

$$\boldsymbol{P(i_1) = sigmoid(\beta_1 \cdot BI(i_1) + \beta_0)}$$

(式 2.4)

ターゲット行動が生じる確率 ($P(a_1)$) は、$P(i_1)$ と（信念ではなく実際の） $P(a_1|i_1)$ の関数である。ターゲット行動をとろうと意図した際に成功裏にターゲット行動をとる実際の確率を実際の自己効力と呼び、$P_{actual}(a_1|i_1)$ と書くことにする。しかしながら、実際の自己効力は計測が難しいため、しばしば知覚された自己効力感を代用品としてターゲット行動が生じる確率が見積もられることが多い。

$$\boldsymbol{P(a_1) = P_{actual}(a_1|i_1) \cdot P(i_1) \fallingdotseq P(a_1|i_1) \cdot P(i_1)}$$

(式 2.5)

$P(a_1)$ が得られると、研究者は行動変容研究で予測したい行動の生成確率を計算することができる。よって、典型的な計画的行動理論の質問紙には、$P(s_n|a_1)$, $U_{self}(s_n)$, $U_k(a_1)$, m_k そして $P(a_1|i_1)$ についての質問が含まれている。

3. 決定理論的な行動変容モデル

上で述べたように、既存の行動変容モデルの精度を上げるために、行動経済学のモデルと組み合わせる試み (Roberto and Kawachi 2015) は、既存の行動変容モデルが期待効用理論と不整合であるために難しいものとなっている。

ここでわれわれは期待効用理論と整合的な新しい行動変容モデルである、決定理論的な行動変容モデルを提案する。決定理論的な行動変容モデルでは、期待効用理論に主観的規範と自己効力感を組み込む。そのために期待効用理論でも存在していた、状態集合 ($S = \{s_1, s_2, \ldots, s_n, \ldots, s_N\}$)、行動集合 ($A = \{a_1$: ターゲット行動を実行する, a_2: ターゲット行動を実行しない$\}$) に加えて、意図集合 ($I = \{i_1$: ターゲット行動をとろうと意図する, i_2: ターゲット行動をとろうと意図しない$\}$) を考える (図 2.1C)。

決定理論的な行動変容モデルでは、i_h ($h = 1, 2$) の生成はそれらの期待効用 ($E[U_{total}|i_h]$) によって決まる。$E[U_{total}|i_h]$ は、意図 i_h が与えられた際にどのような結果が生じるかに関する主観確率 ($P(s_n|i_h) = \Sigma_{j=1}^{2} P(s_n|a_j) * P(a_j|i_h)$) と全体の効用 ($U_{total}$) によって決まるものとする。全体の効用は、定義域を行動と状態とする、エージェント自身の効用と他者の効用の重みつき和によって決まるものとする ($U_{total} = U_{self} + wU_{others}$)。よって、期待効用 $E[U_{total}|i_h]$ は、

$$\boldsymbol{E[U_{total}|i_h] = \sum_{j=1}^{2} P(a_j|i_h) \sum_{n=1}^{N} P(s_n|a_j)\{U_{self}(a_j, s_n) + wU_{others}(a_j, s_n)\}}$$

(式 2.6)

となる。計画的行動理論と比較するために 式 2.6 を書き換えると、

$$\boldsymbol{E[U_{total}|i_h] = P(a_1|i_h)\sum_{n=1}^{N} P(s_n|a_1)\{U_{self}(a_1,s_n) + wU_{others}(a_1,s_n)\}}$$
$$\boldsymbol{+ P(a_2|i_h)\sum_{n=1}^{N} P(s_n|a_2)\{U_{self}(a_2,s_n) + wU_{others}(a_2,s_n)\}}$$

(式 2.6')

のようになる。

計画的行動理論の 式 2.3' と決定理論的な行動変容モデルの式 2.6' は以下の 5 つの点で異なっている (図 2.1B と 2.1C)：

(1) $E[U_{total}|i_1]$ は期待効用の形になっている。$E[U_{total}|i_1]$ は通常の期待効用理論の $E[U_{self}|a_1]$ に主観的効用と自己効力感を加えた自然な拡張になっている。一方、計画的行動理論の $BI(i_1)$ は期待効用とみなすことはできない。

(2) 提案モデルは、i_1 が与えられた際の期待効用 ($E[U_{total}|i_1]$) のみではなく、i_2 が与えられた際の期待効用 ($E[U_{total}|i_2]$) も明示的に考えている。一方、計画的行動理論は i_1 のときの行動意図 ($BI(i_1)$) しか明示的に考えていない。この違いは、$P(i_1)$ と $P(a_1)$ を考える際に重要な違いとなる。

(3) 提案モデルの $U_{self}(a_j, s_n)$ と $U_{others}(a_j, s_n)$ は計画的効用理論の $U_{self}(s_n)$ と $U_{others}(a_1)$ よりも柔軟性が高い。計画的行動理論では、エージェントの行動コストに対するエージェント自身の効用や、行動の結果に対する他者の効用を考えることができなかった。

(4) 提案モデルの $E[U_{total}|i_1]$ はエージェントがターゲット行動をとろうとしたが成功裏に実行できなかった場合を考慮に入れているが、計画的行動理論の $BI(i_1)$ はそのような状況を考慮に入れていない。

(5) 知覚された自己効力感が、行動が与えられた際の期待効用に対して、提案モデルでは掛け算されているが、計画的行動理論では足されている。

期待効用理論のように、意図選択ルールをシグモイド関数で表現すると (Luce 1959; Sutton and Barto 1998)、

$$\boldsymbol{P(i_1) = sigmoid(\beta_1 \cdot \{E[U_{total}|i_1] - E[U_{total}|i_2]\} + \beta_0)}$$

(式 2.7)

となる。式 2.4 と式 2.7 の違いは、式 2.7 においては、$E[U_{total}|i_2]$ が明示的に考慮されているのに対して、式 2.4 ではそうではないことである。$E[U_{total}|i_2]$ が全員同じ場合には、定数項で暗黙的に $E[U_{total}|i_2]$ を表現できるが、人によって異なる場合には $P(i_1)$ の計算結果が異な

ることになる。

ターゲット行動の生成確率は

$$\boldsymbol{P(a_1) = P_{actual}(a_1|i_1) \cdot P(i_1) + P_{actual}(a_1|i_2) \cdot P(i_2) \fallingdotseq P(a_1|i_1) \cdot P(i_1) + P(a_1|i_2) \cdot P(i_2)}$$

(式 2.8)

と書くことができる。式 2.5 と式 2.8 の違いは、式 2.8 はターゲット行動をとろうとしなかったにもかかわらずターゲット行動が生じる場合を明示的に考慮に入れているのに対して、式 2.5 ではそうではないことである。$P_{actual}(a_1|i_2)$ か $P(i_2)$ が 0 であるとき以外はこの違いにより、式 2.5 と式 2.8 の計算結果が異なったものになる。提案モデルでは式 2.6 から 式 2.8 によってターゲット行動の生成確率を計算できる。

まとめると、提案モデルは行動変容のための期待効用理論の自然な拡張になっている。

4. 提案モデルの計画的行動理論に対する優位性

この節では、前節の式 2.3’ と式 2.6’ の 5 つの相違点のうち最後のものに焦点を当てて、提案モデルの計画的行動理論に対する優位性を主張する。上述のように、知覚された自己効力感は行動が与えられた際の期待効用対して、提案モデルでは掛け算されているが、計画的行動理論では足されている。

ここでは例として、固いビンのふたを開けるケースを考えてみよう。単純のために、エージェント以外の個人がいない状況を考える。ターゲット行動 (a_1) は「ビンのふたを開けるのに十分な力で手首をひねる」である。それに対応して、i_1 は「ビンのふたを開けるのに十分な力で手首をひねろうとする」であり、s_1 は「ビンのふたが開く」、s_2 は「ビンのふたが開かない」である。

計画的行動理論だと、行動意図は、(1) 行動への態度：これはビンの中身がエージェントにとってどの程度の価値を持つかによって決まる、(2) 主観的規範：今は他者がいない状況を考えているのでこれは無視できる、(3) 知覚された自己効力感：これは手首に力を入れようとしたときに、どのくらいの確率でビンが開くくらい強い力を出せるかに関する信念、の 3 つの要因によって決まる。このケースでは 3 つの要因の重み係数は全て正の値になるものと考えられる。ここで、エージェントが脊髄を損傷してしまい、全身麻痺になってしまった場合を考えよう。このとき、自己効力感は 0 になるものと考えられる。一方、行動への態度（と主観的規範）は変わらない。計画的行動理論の行動意図は、3 つの要因の重みつき和によって決まる。そのため、計画的行動理論によると、手を動かす能力の有無によらず、ビンの中身の価値に応じて手首をひねろうという行動意図が生まれるということになる。この予測は非現実的であり、計画的行動理論に対する反例とみなすことができる。

一方、提案モデルでは行動への態度に対して、自己効力感 (= 0) が掛け算されているため、

手首に力を入れようとする行動意図が生まれないことをきちんと予測することができる。この例は、提案モデルの優位性を示す一例とみなすことができる。

5. 議論

本章では計画的行動理論を、主観的規範と自己効力感を組み込むことで通常の期待効用理論を改良しようとしたものとみなせることを示した。計画的行動理論は比較的シンプルであり、3 つの要因が実際に行動変容に効いてくるため、よく使われて大きな成功を収めた (Sheeran, Maki et al. 2016)。 多くの人々が計画的行動理論を用いることにより、人間の社会行動を理解、予測することに成功してきた (Van Lange, Kruglanski et al. 2011) 。計画的行動理論を用いた介入によって 2/3 の研究で実際に人々の行動が変容したという報告もある (Hardeman, Johnston et al. 2002)。

しかしながら、計画的行動理論は重大な問題を抱えている。主観的規範と自己効力感が通常の期待効用理論の要素に単純に足されてしまっているため、もはや期待効用理論と整合的ではなく、それゆえに行動経済学的なモデルと組み合わせて使うことができなくなってしまっている。この問題点を解決するために、本章では主観的規範と自己効力感を期待効用理論に整合的な形で組み込んだ新しいモデルである、決定理論的な行動変容モデルを提案した。

提案モデルは期待効用理論と整合的であるため、簡単に拡張することが可能である。第1 に、時間割引を導入することにより異時点間意思決定の問題を扱うことができる。特に、行動経済学でよく用いられる双曲割引を用いることで先延ばしを表現することができる (Story, Vlaev et al. 2014)。第 2 に、マルコフ決定過程の枠組みを用いることで、多状態モデルに拡張することができる。多状態モデルは、病気の状態と健康な状態の両方で意思決定するような場合に用いることができる (Sonnenberg and Beck 1993; Sutton and Barto 1998)。脳の報酬系はマルコフ決定過程の枠組みを使って分析されることが多いため、この方向性の拡張により、薬物中毒のような異常な行動についての薬理学的なモデルと行動変容モデルを組み合わせることが可能となる (Redish 2004; Rangel, Camerer et al. 2008)。第 3 に、本章では U_{total} を U_{self} と U_{others} の和によって定義したが、社会的選好に関する研究を参照することでもっと他の定義を用いることができるものと思われる (Fehr and Krajbich 2014)。 第 4 に、提案モデルは道徳性のモデルに用いることができる (Crockett 2013)。提案モデルは、期待効用理論に意図と行動の区別を持ち込んだものになっており、この区別は道徳判断の重要な特徴となっている (Cushman 2008)。提案モデルの効用関数は道徳的価値を表現するのに適している。

決定理論的な行動変容モデルを用いることで、既存の行動変容研究と行動経済学のよりよいコンビネーションが生まれれば幸甚である。

第 3 章 ベイズ推定のバイアスのモデル化

意思決定と認知制御は不確実性のある非定常的な環境で適応的に行動するために必要不可欠なものである (Fuster, 2015)。心理学、神経科学、工学などのさまざまな分野の研究は、意思決定と認知制御は異なった脳部位に実装されていることを示してきた (Fuster, 2015; Stuss and Knight, 2013)。しかし、これらのプロセスの計算論原理は議論のさなかにある (Gazzaniga 2009; Friston 2010)。

最近の理論的な研究が提案するところによると、意思決定や認知制御を含むさまざまな認知機能をベイズ推定によって説明することが可能である (Friston 2010; Beck, Ma et al. 2012; Pouget, Beck et al. 2013)。この考え方は、運動制御や知覚的意思決定において、ほぼベイズ最適な方法で情報が統合されているという多くの実験的な研究によって支持されている (Wolpert, Ghahramani et al. 1995; Ernst and Banks 2002; Körding and Wolpert 2004)。

Körding and Wolpert (2004) は到達課題において事前に学習された確率分布と新しく追加された確率分布の情報をベイズ的に組み合わせていることを示した (Körding and Wolpert 2004)。また、Brunton et al. (2013) は知覚的意思決定が、知覚のノイズの影響は受けるが、情報蓄積過程のノイズの影響は受けていないことを示した (Brunton, Botvinick et al. 2013)。情報蓄積過程はベイズ推定と見なすことができるので (Bitzer,

Park et al. 2014)、彼らの発見は意思決定における情報の蓄積がベイズ最適な形で行われていることを示唆している。

一方、知覚的意思決定の蓄積モデルとして知られる Decision Field Theory (DFT) や Leaky Competing Accumulator model (LCA) (Busemeyer and Townsend 1993; Usher and McClelland 2001) は、意思決定のバイアスである魅力効果などの文脈効果や primacy/recency 効果を説明するために、情報蓄積過程にもバイアスを想定している。これらのモデルはベイズ最適な推定からの乖離を考慮に入れていると見なすことができることを後に示す。

古典的な経済的意思決定の研究では、人間は最適に行動すると想定されている (von Neumann and Morgenstern 1947)。しかしながら、確率判断における認知バイアスの存在が繰り返し示されてきた (Kahneman, Slovic et al. 1982)。よく知られた問いとして以下のようなものがある。「ある集団の中に A 病の人が 1/10,000 の割合で存在する。 病気だった場合 99%の確率で陽性になり、健康だった場合 99%で陰性になる検査を受けたところ、病気という結果が出た人が、実際に病気である確率はどれくらいだろうか」。多くの人は正しい値である約 1%よりも大きな値を答える傾向にあることが知られており、「基準率の無視のバイアス」と呼ばれている (Kahneman, Slovic et al. 1982)。他のバイアスである、代表性バイアスや保守性バイアス、アンカリングと調整もまた、最適なベイズ推定からの乖離と見なすことができる (Kahneman, Slovic et al. 1982)。

知覚的意思決定、経済的意思決定、確率判断等のさまざまな心理的な現象は最適なベイズ推定もしくはそこからのシステマティックな乖離として記述されてきた。しかしながら、知覚的意思決定と経済的意思決定を統一的な観点から論ずる試みはこれまであまりなされてこなかった (Summerfield and Tsetsos 2012)。

本章では、指数バイアスを含むように拡張されたベイズ推定を考えることで、意思決定や確率判断のバイアスを説明できることを示す。また、指数バイアス付きベイズ推定の神経モデルへの実装についても議論する。はじめに、二つの指数バイアスを含んだベイズ推定について記述する。次に、最尤推定、ベイズ推定、MAP 推定の三つをバイアス付きベイズ推定の二つのバイアスパラメータから成る二次元空間上に位置づけられることをしめす。このパラメータ空間上に行動経済学等の研究で見つかった行動バイアスも位置づけられることも示す。さらに、神経結合のゲイン調整を考慮に入れることで、バイアス付きベイズ推定の神経モデルを簡単に作ることができることを示す (Goldman, Compte et al. 2009)。これは、意思決定の神経モデルである LCA (Usher and McClelland 2001) と確率的ポピュレーションコーディング (Probabilistic Population Codes, PPC) (Ma, Beck et al. 2006) を組み合わせたものだと見なすことができる。最後に、この枠組みと認知制御の関係について議論する。

1. バイアス付きベイズ推定

ベイズ推定に基づく手法はデータ数が限られている際の推論と意思決定においてとても強い力を発揮することができる。この方法は、遺伝学、言語学、イメージプロセシング、宇宙学、生態学、機械学習、心理学、神経科学など多くの分野に用いられてきている (Stone 2013)。

これらすべてのケースで、ベイズ的手法における解法はとても簡単な公式であるベイズの定理に基づいている。ベイズの定理によると、データ D を観察した後の仮説 H の確率は、データの尤度と事前の信念の積に比例する ($P(H|D) \propto P(D|H) * P(H)$)。

しかしながら、人やその他の動物の行動を見ると最適なベイズ推定からの乖離があることが繰り返し実験的に示されてきた (Kahneman, Slovic et al. 1982)。いくつかのケースでは、そのような乖離は指数バイアスを導入することによって説明されてきた (Nassar, Wilson et al. 2010; Soltani and Wang 2010; Payzan-LeNestour and Bossaerts 2011; Payzan-LeNestour and Bossaerts 2012; Payzan-LeNestour, Dunne et al. 2013)。主にそれは、バイアスのレベルを回帰によって求められるからであった。事前確率と尤度に関するバイアスはそれぞれ独立に議論することができるが、ここで、事前確率と尤度について同時に指数バイアスを入れることを考える (図 3.1)。

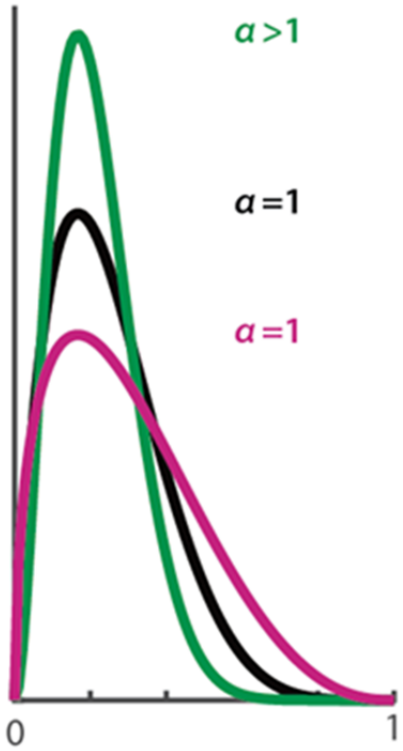

図 3.1 指数バイアス ($P \propto P^{\alpha}, 0 \leq \alpha$)。β 分布が例として示されている。元の分布 (黒) が $\alpha < 1$ のときフラットになっている (マゼンタ) 、$\alpha > 1$ のときシャープになっている (緑)。

$$P(H|D) \propto P(D|H)^{\beta} P(H)^{\alpha}$$

(式 3.1)

$$\log\big(P(H|D)\big) = \beta \log\big(P(D|H)\big) + \alpha \log\big(P(H)\big) + \text{const.}$$

(式 3.1')

ここで、$\propto$ は比例、α は事前確率の重み、β は尤度の重み、そして *const.* は定数を意味する。 式 3.1' は式 3.1 の対数をとったものである。いくつかの研究では、ベイズの定理を対数尤度比と対数オッズ比の和として表現している (Grether 1980; Soltani and Wang 2010)。しかしながら、これらの研究では選択肢が二つの場合のみを考えており、二つ以上の選択肢の場合において問題となる選好逆転などの現象を扱えないという問題点がある (Churchland and Ditterich 2012)。

ここで、バイアスの強さ (α, β) は正であると仮定する。α = β = 1 のとき、推定は普通の最適なベイズ推定になっている。事前確率の重みが普通より小さい場合 (0 < α < 1)、事前分布は元の分布よりフラットになる。このケースでは、最初の事前分布の影響が、推定ごとにどんどん小さくなっていく。このタイプの推論バイアスは昔の影響がどんどんなくなっていくことから *forgetting* と呼ばれている (Peterka 1981; Nassar, Wilson et al. 2010; Payzan-LeNestour and Bossaerts 2011; Payzan-LeNestour and Bossaerts 2012)。非定常的な環境の場合、古い情報は新しい情報に比べて重要性が低い。そのため、環境についての完全な情報を我々が持っていない場合に、 *forgetting* を用いることで環境の

非定常性を考慮に入れることができる (Peterka 1981; Kulhavý and Zarrop 1993)。

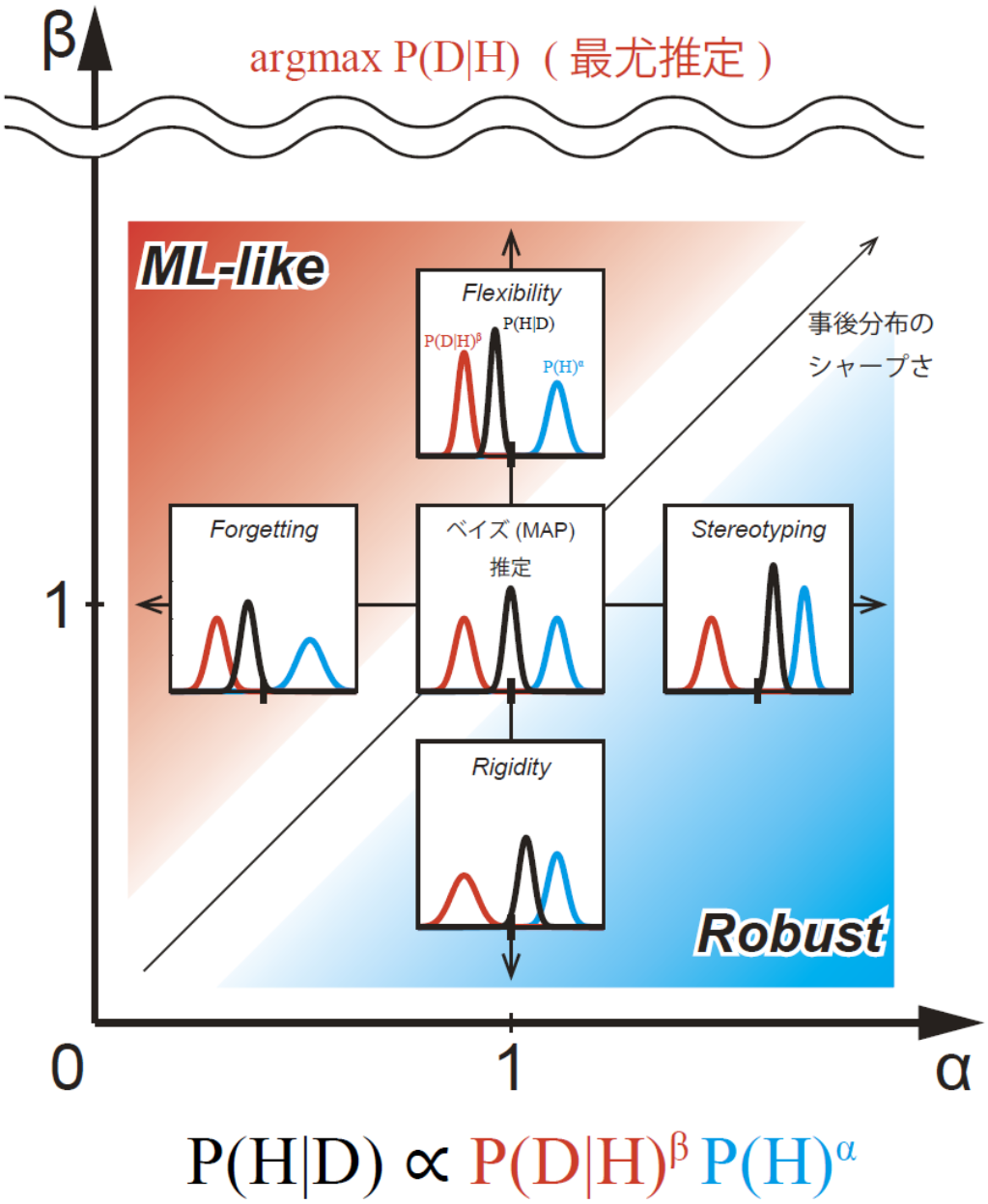

図 3.2 指数バイアス付きベイズ推定 ($P(H|D) \propto P(D|H)^{\beta} * P(H)^{\alpha}$) のバイアス平面。事前確率（青）と尤度（赤）を異なった重みづけで組み合わせることで異なった事後確率（黒）が生成される。

事前分布の重みが通常よりも大きい場合 ($\alpha > 1$) 事前分布は元よりもシャープなものになる (図 3.2a)。この場合、事前分布の影響は通常よりも強くなる。昔の情報にどんどん凝り固まっていくため、このタイプのバイアスを *stereotyping* と呼ぶことにする。

尤度の重みが通常よりも弱い場合 ($0 < \beta < 1$)、尤度は元よりもフラットになる。この場合、事後分布への尤度の影響は弱くなっており、観察されたデータをあまり考慮に入れてないことになるので、このようなバイアスを *rigidity* と呼ぶことにする。もしも観察されたデータが、外れ値だった場合、*rigidity* によって外れ値の影響を減らすことができる (Agostinelli and Greco 2012; Agostinelli and Greco 2013)。

尤度の影響が通常よりも大きい場合 ($\beta > 1$)、尤度は元よりもシャープになる。この場合、観察されたデータが事後分布に通常より強い影響を及ぼす ので *flexibility* と呼ぶことにする。β が $+\infty$ へと発散するとき、事後分布は尤度の最頻値のみの値をとるようになる。そのため、*flexibility* はベイズ推定と最尤推定の中間的な推定方法を提供していると見なすことが可能である (図 3.2)。

2. バイアス付きベイズ推定に基づくパラメータ推定

前の節では、指数バイアス付きベイズ推定と導入した。ここでは、バイアス付きベイズ推定のバイアス平面（α-β 平面）上にさまざまなパラメータ推定法を位置づける (図 3.2)。

事前分布と尤度が与えられているとき、バイアス平面上の各点がそれぞれ異なる事後分布に対応している。そのため、各点は異なった推定方法に対応していると見なすことができる。また、事後分布そのものを推定値として返すこともできるが、平均値や最頻値や中央値を返すこともできる。

もしも、事後分布の最頻値を返す場合、$(\alpha, \beta) = (1, 1)$ は通常の MAP 推定に対応している。$\alpha < \beta$ のとき、尤度の方が事前分布よりも影響が強いので、この領域の推定方法を *ML-like* 推定と呼ぶことにする。この領域は MAP 推定と最尤推定の中間的な推定方法と考えることができる。特に、$\alpha = k_1$ かつ $\beta = 1$ の時、$k_1 = 0$ ならば最尤推定に一致して、$k_1 = 1$ ならば MAP 推定に一致する。$0 < k_1 < 1$ のとき両者の中間的な推定方法になっている。さらに、$\alpha = 1$ かつ $k_2 = 1/\beta$ の時、$k_2 = 0$ ならば最尤推定に一致して、$k_2 = 1$ ならば MAP 推定に一致している。さらに、$0 < k_2 < 1$ の時、もう一つ別のタイプの中間的な推定方法を考えることができる。

一方、$\alpha > \beta$ のとき、尤度の方が事前分布よりも影響が弱いので、この領域の推定方法を *robust* な推定と呼ぶことにする。この領域においては、MAP 推定と「最大事前確率推定」の中間的な推定方法を考えることができる。$\alpha = k_1$ かつ $\beta = 1$ の時、$k_1 = 1$ ならば MAP 推定に一致して、$k_1 \to +\infty$ で最大事前確率推定に一致する。$k_1 > 1$ で両者の中間的な推定方法を考えることができる。さらに、$\alpha = 1$ かつ $k_2 = 1/\beta$ の時、$k_2 = 1$ ならば、MAP 推定に一致して、$k_2 \to +\infty$ で最大事前確率推定に一致する。また、$k_2 > 1$ で

別のタイプの両者の中間的な推定方法を考えることができる。

事後分布そのものを返す場合 $(\alpha, \beta) = (1, 1)$ は通常のベイズ推定に対応している。*ML-like* の領域において、$\alpha = 1$ かつ $k_2 = 1/\beta$ の時、$k_2 = 0$ ならば最尤推定に一致して、$k_2 = 1$ ならばベイズ推定に一致する。$0 < k_2 < 1$ で両者の中間的な推定方法になっている。*Robust* な領域において、$\alpha = 1$ かつ $k_2 = 1/\beta$ の時、$k_2 = 1$ ならばベイズ推定に一致して、$k_2 \to +\infty$ ならば「事前確率推定」に一致する。また $k_2 > 1$ で両者の中間的なものを考えることができる。

確率判断のバイアスをバイアス平面上に位置づける

この節では、複数の認知バイアスをバイアス平面上に位置づける (図 3.2)。人間の確率判断の文脈において、基準率の無視、代表性バイアス、保守性バイアス、アンカリングと調整といった、さまざまなベイズ推定からのシステマティックな乖離が知られている (Tversky and Kahneman 1974; Kahneman, Slovic et al. 1982)。バイアス平面において、基準率の無視と代表性バイアスは尤度の影響が事前分布に比べて強いため *ML-like* 推定の領域 $(\alpha < \beta)$ に対応している。ここで、このバイアスは *forgetting* $(\alpha < 1)$ によっても *flexibility* $(\beta > 1)$ によっても生じることを強調しておく。これまでこの二つははっきり区別されてこなかったが、事前確率の影響が弱くなっているからバイアスが生じているのか、尤度の影響が強くなっているからバイアスが生じているかの区別はバイ

アスのメカニズムを考えるうえで重要なものだと考えられる (Brock 2012; Pellicano and Burr 2012)。

保守性バイアスとアンカリングと調整は *robustness* ($\alpha > \beta$) に対応していると見なすことができる。事後確率が通常より事前確率に近いこのような推定は *stereotyping* ($\alpha > 1$) によっても *suspiciousness* ($\beta < 1$) によっても生じるが、これらは別のメカニズムであることに関しても強調しておく。これまでの研究のほとんどは選択肢が２つの場合のみを扱っていたが (Grether 1980; Soltani and Wang 2010)、 ここでのバイアス平面によるアプローチは選択肢が 2 つ以上の場合にも用いることができる。また、これらの心理バイアスは適応的な意味があるかもしれないことが提案されている (Gigerenzer and Goldstein 1996)。

3. バイアス付きベイズ推定の神経モデル

確率分布の神経表象

いくつかの研究において、ベイズ推定をおこなう神経回路の計算論モデルが提案されてきている (Jazayeri and Movshon 2006; Ma, Beck et al. 2006)。ベイズ推定をおこなうためには神経回路が確率分布を表象する必要がある。通常、シナプス入力は加法的に機能すると考えられているため (Kandel, Schwartz et al. 2000)、ベイズ推定は神経活動の和によって実現されると考えられる。足し算によってベイズ推定をおこなうことを可

能にするためには神経活動は確率の対数を表現する必要がある。実際、いくつかの実験によって神経細胞が確率の対数を表現していることが示されてきた (Yang and Shadlen 2007; Kira, Yang et al. 2015)。

二つの理論的モデルが、神経細胞集団が対数確率をエンコードしていることを提唱してきた (Pouget, Beck et al. 2013; Ma and Jazayeri 2014)。最初のモデルは対数尤度コーディングと呼ばれるもので、神経細胞の発火率が確率の対数に比例していると主張するものである (Barlow 1969; Jazayeri and Movshon 2006; Pouget, Beck et al. 2013)。たくさんの神経細胞を用いることでこの方法によって確率分布を表現することが可能である (Yang and Shadlen 2007; Pouget, Beck et al. 2013)。

もう一つのより洗練されたモデルは確率的ポピュレーションコーディングと呼ばれるものである。このモデルは対数尤度コーディングと基底関数を組み合わせて、logP(x) が以下のような式で表現できると考える (Ma, Beck et al. 2006; Pouget, Beck et al. 2013):

$$\log \mathrm{P}(\mathrm{x}) \;=\; \sum_i \mathrm{r}_i \mathrm{h}_i(\mathrm{x}) \;+\; \mathrm{const} \;=\; \mathbf{r} \cdot \mathbf{h} \;+\; \mathrm{const.}$$

(式 3.2)

ここで、r_i は i 番目の神経細胞の発火率、$\mathbf{r} = (r_1, r_2, \dots, r_n)^T$、$h_i(x)$ は i 番目の神経細胞の基底関数、$\mathbf{h} = (h_1(x), h_2(x), \dots, h_n(x))^T$ である。

この論文では、確率的ポピュレーションコーディングに基づくバイアス付きベイズ

推定の神経モデルを提案する。

神経細胞集団におけるバイアス付きベイズ推定の実装

以下で、通常のベイズ推定とバイアス付きベイズ推定をおこなう神経モデルを考えるうえで、神経細胞集団の層を二つ考える。最初の層は事前確率を表象する神経細胞集団と、尤度を表現する神経細胞集団からなる。二番目の層は事後分布を表象する神経細胞集団からなる (図 3.3a)。

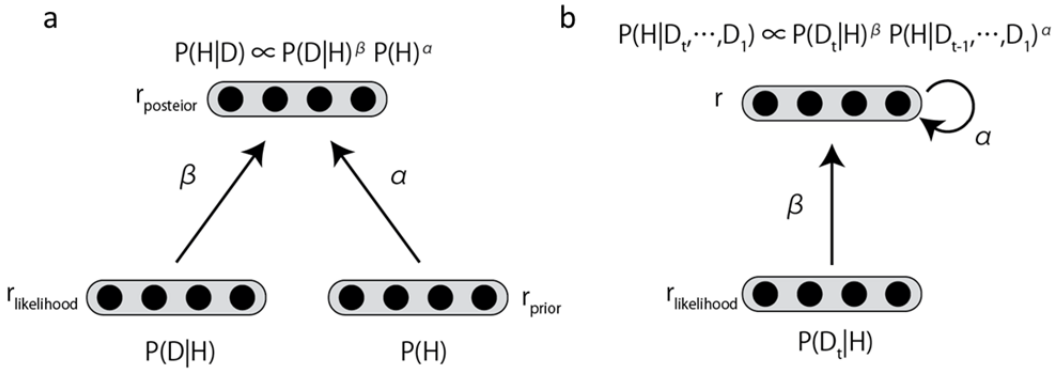

図 3.3 バイアス付きベイズ推定の神経モデル。 (a) フィードフォワードなモデル。最初の層は事前確率と尤度をそれぞれコードする。次の層でそれらが足しあわされることで事後確率が計算される。層の間の結合のゲインがバイアスの強さに対応している。(b) 再帰的結合によるモデル。最初の層は外部入力である尤度をコードしている。次の層は事後確率（= 次の時間の事前確率）をコードしている。

対数確率の和は事前確率と尤度の積に等しいため (Ma, Beck et al. 2006)、確率的ポピュレーションコーディングにおいて通常のベイズ推定は、事前確率を表象する神経細胞集団と尤度を表現する神経細胞集団の活動を事後確率を表象する神経細胞集団において足し算することによって実現される。

ここで、指数バイアスを含んだベイズ推定 ($P(H|D) \propto P(D|H)^{\beta} * P(H)^{\alpha}$) の神経モデルを提案する。バイアス付きベイズ推定は、通常のベイズ推定をおこなう神経モデルにおいて、事後確率を表象する層への入力のゲイン (α, β) を変えることによって簡単に実装することが可能である (式 3.3) (図 3.3a)。

$$\log(P(D|H)^{\beta}P(H)^{\alpha}) = \beta \log P(D|H) + \alpha \log P(H) = \beta\, \mathbf{r}_{\text{likelihood}} \cdot \mathbf{h} + \alpha\, \mathbf{r}_{\text{prior}} \cdot \mathbf{h} + \text{const} = (\beta\, \mathbf{r}_{\text{likelihood}} + \alpha\, \mathbf{r}_{\text{prior}}) \cdot \mathbf{h} + \text{const} = \mathbf{r}_{\text{posterior}} \cdot \mathbf{h} + \text{const}$$

(式 3.3)

ここで、$\mathbf{r}_{\text{prior}}$、$\mathbf{r}_{\text{likelihood}}$、$\mathbf{r}_{\text{posterior}}$ は事前確率、尤度、事後確率を表象する層の神経細胞の発火率である。α と β はそれぞれ、事前確率を表象する層から事後確率を表象する層への投射のゲイン、尤度を表象する層から事後確率を表象する層への投射のゲインである。そのため、$\mathbf{r}_{\text{posterior}}$ は $\beta\, \mathbf{r}_{\text{likelihood}} + \alpha\, \mathbf{r}_{\text{prior}}$ によって計算することができる。

再帰的な結合と神経積分器

ベイズ更新において、事後分布は次の時間ステップの事前分布になることが多い。

ベイズ更新の時間ステップは課題によって異なっており、試行ベースであったり (Glimcher 2003)、より細かい試行内での更新を考えることもある (Bogacz, Brown et al. 2006; Kira, Yang et al. 2015)。このような繰り返し更新は、時間ステップの細かさによらず再帰的な神経回路によって実現することができる (Goldman, Compte et al. 2009; Bitzer, Park et al. 2014; Kira, Yang et al. 2015; Chandrasekaran 2017)。 神経積分器モデルは発火率を再帰的な入力によって保存する仕組みとして考えられてきた (Goldman, Compte et al. 2009)。神経積分器のダイナミクスは発火率方程式で表現できるため (Dayan and Abbott 2001; Goldman, Compte et al. 2009)、バイアス付きベイズ推定は式 3.4 によって表現できる (図 3.3b)。

$$\tau_{\mathrm{neuron}}\, \mathrm{d}\mathbf{r}/\mathrm{dt} = -\mathbf{r} + \mathrm{Input} = -\mathbf{r} + \alpha * \mathbf{r} + \beta * \mathbf{r}_{\mathrm{likelihood}}(\mathrm{t})$$

(式 3.4)

ここで、τ_{neuron} は神経細胞の時定数、 $\boldsymbol{r}$ は事前確率と事後確率を表現する細胞集団の発火率、*Input* は事前確率と事後確率を表象する層への入力 ($\alpha\,\mathbf{r} + \beta\,\mathbf{r}_{\mathrm{likelihood}}(\mathrm{t})$)、$\alpha$ は再帰的結合の強さ、β は前向き結合の強さ、$\boldsymbol{r}_{likelihood}$ は尤度を表現する神経細胞集団の発火率である。

神経積分器モデルにおいて、再帰的結合の強さは神経細胞が発火率を保存するために必須である (Goldman, Compte et al. 2009)。外部入力なしで発火率を保持できる場合、再帰的な入力と発火率の減衰のバランスが取れているため、神経積分器は *balanced* (α

= 1) であると言われる。もし、再帰的結合が弱くて発火率が減衰していくとき、神経積分器は *leaky* ($\alpha < 1$) であると言われる。この場合、発火率とそれによって表現される確率分布は時間とともに減衰していく (*forgetting*)。もし、再帰的結合が強くて、どんどん大きくなっていくとき神経積分器は *unstable* ($\alpha > 1$) であると言われる。この場合、発火率とそれによって表現される確率分布は時間とともにシャープになっていく (*stereotyping*)。

バイアス付きベイズ推定と 2 肢強制選択法

2 肢強制選択における意思決定は最も単純な意思決定の一つであるため、とてもよく研究されてきている (Stone 1960; Ratcliff 1978; Busemeyer and Townsend 1993; Gold and Shadlen 2001; Roe, Busemeyer et al. 2001; Shadlen and Newsome 2001; Usher and McClelland 2001; Wang 2002; Mazurek, Roitman et al. 2003; Usher and McClelland 2004; Bogacz, Brown et al. 2006; Wong and Wang 2006; Kiani, Hanks et al. 2008; Tsetsos, Gao et al. 2012)。Drift-diffusion model (DDM) (Ratcliff 1978) や DFT (Busemeyer and Townsend 1993; Roe, Busemeyer et al. 2001)、LCA (Usher and McClelland 2001; Usher and McClelland 2004) といった 2 肢強制選択における蓄積モデルは以下の 3 つの仮定を置いている:

i.) 1 試行の内で各選択肢を支持する証拠が蓄積していく

ii.) このプロセスはランダム変動に従う

iii.) ある選択肢について十分証拠が蓄積した時に決定がなされる

(Bogacz, Brown et al. 2006)。

これらのモデルはサルの lateral intraparietal cortex (LIP) の神経活動に基づいて作られている (Kiani, Hanks et al. 2008; Brunton, Botvinick et al. 2013)。DDM は最適な積分を仮定しており、ベイズ最適な意思決定に対応している (Neyman and Pearson 1933; Wald and Wolfowitz 1948; Wald 1973; Bogacz, Brown et al. 2006; Kira, Yang et al. 2015)。 瞬間的な変動は尤度であると見なせるため、蓄積のスタート地点を事前確率と見なせば、DDM の瞬間的な変動はベイズ推定であると見なすことができる (Bogacz, Brown et al. 2006; Bitzer, Park et al. 2014)。

DDM とは対照的に、primacy/recency 効果や選好逆転効果を説明するために、DFT と LCA は leaky な神経積分器を仮定している (Busemeyer and Townsend 1993; Roe, Busemeyer et al. 2001; Usher and McClelland 2001; Usher and McClelland 2004; Bogacz, Brown et al. 2006)。しかしながら、DFT や LCA をベイズ推定の観点から解釈する試みはこれまでなされてこなかった。DFT や LCA を単純化したものである Ornstein–Uhlenbeck 過程 (Bogacz, Brown et al. 2006) は、再帰的な神経モデルに一致している (図 3.3b)。そのため、DFT と LCA はバイアス付きベイズ推定と見なすことが可能である (図 3.4)。ここで、バイアス付きベイズ推定の枠組みは 2 肢強制選択における意思決定の記述的なモデルと規範的なベイズの観点からの見方を組み合わせるうえで重要であると考えられる (Beck, Ma et al. 2012)。

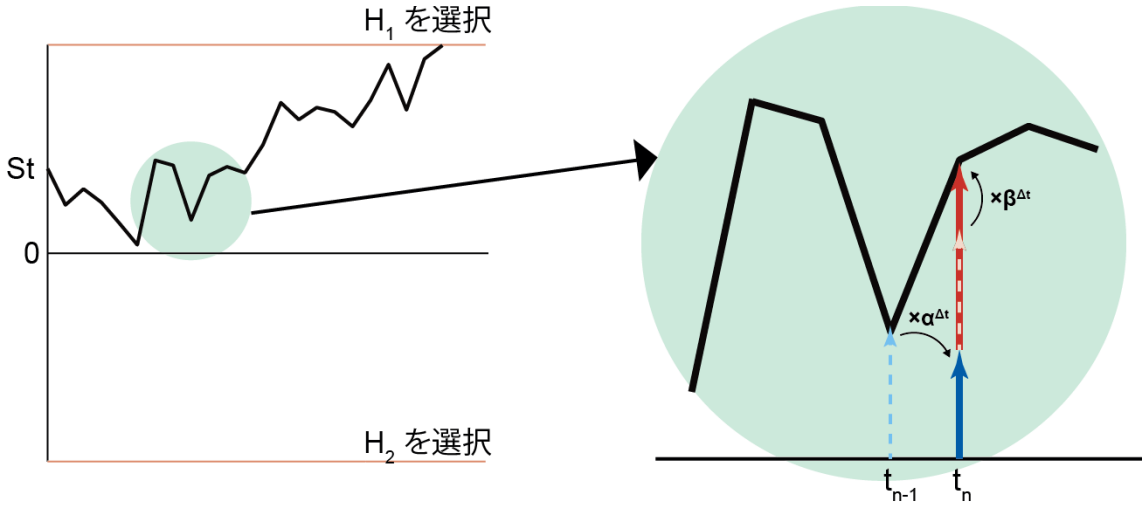

図 3.4 2 肢強制選択課題の蓄積モデルにおけるバイアス付きベイズ推定。二つの仮説 (H_1, H_2) の間での選択のプロセスの例。左図の縦軸は仮説の対数確率 ($\log[P(H_1|D)/P(H_2|D)]$) であり、横軸は時間。

ベイズ推定に加えて、内的な効用関数も行動の選択において考慮に入れられるべきである (Ernst and Bulthoff 2004)。各選択肢にそれぞれ異なった価値が割り振られているとき、効用関数 U(x) も行動選択にとって重要である。 P(x) が発火率 $\mathbf{r}_{prior/posterior}$ によって表象されていたように、U(x) もまた、細胞集団の発火率 $\mathbf{r}_{reward}$ によって表象することができる。期待効用 EU(x) は確率と効用の積によって計算できるため、ベイズ推定と効用を組み合わせて期待効用を表現することは $\mathbf{r}_{prior/posterior} + \mathbf{r}_{reward}$ (log(EU(x)) = log(P(x)) + log(U(x)) によって実現することができる。実際、最近の研究によると選択肢間での報酬量の違いは、尤度に対応する変動幅ではなく、事前確率に対応する蓄積過程のスタート地点に影響していることが示されている (Rorie, Gao et al. 2010; Summerfield and Koechlin 2010; Mulder, Wagenmakers et al. 2012)。そのため、ベイズ推定と期待効用の計算は同じ神経メカニズムによって実現できるものと考えられる (Friston, Schwartenbeck et al. 2013)。

また、視覚的注意によって蓄積モデルの平均変動幅が変化することが知られており (Krajbich, Armel et al. 2010)、この結果は効用に関しての指数バイアスが存在することを示唆する。

4. 議論

ゲイン調整としての認知制御とバイアス付きベイズ推定

バイアス付きベイズ推定の神経モデルが神経細胞集団のゲイン調整によって説明できることを上で議論した。ここで、ゲイン調整すなわちバイアスの強さの調整が認知制御によって行われている可能性について議論する。

認知制御は 二つのステップによって起きると考えられている。一つ目は anterior cingulate cortex (ACC) が認知制御の必要性をモニターして、制御信号を prefrontal cortex (PFC) に送るステップ。二つ目は PFC がさまざまな皮質領域の活動を調整するトップダウンの認知制御シグナルを送るステップである (Botvinick, Braver et al. 2001; Botvinick, Cohen et al. 2004; Matsumoto 2004; Shenhav, Botvinick et al. 2013)。

- ストループ課題への適用

ストループ課題は認知制御の神経メカニズムを調べるための課題としてよく知られている。この課題は単語読み課題と色の名前課題から成る。単語読み課題では、被験者はある色で書かれた色を意味する単語を読み上げることを要求される。ここで、文字の色と単語の意味する色は同じ時 (*congruent* 条件) と別な時 (*incongruent* 条件) がある。例えば、赤色で書かれた「緑」という文字を見た時に被験者は、「みどり」と答えることを要求される。現代社会の人々は文字を読むことが多いため、この課題には

あまり認知制御は必要とされない。色の名前課題では、被験者はある色で書かれた色を意味する単語について、文字の色を答えることを要求される。この課題でも、文字の色と単語の意味する色は同じ時と別な時がある。例えば赤色で書かれた「緑」という文字を見た時に、「あか」と答えることを要求される。現代の人々は文字の色を読み上げることはあまりないため、この課題は単語読み上げ課題に比べて比較的認知制御が必要な課題であることが知られている。特に、文字の色と単語の意味する色が別な場合 (*incongruent* 条件) の方が、反応時間も間違い率も大きいことが知られており、この効果はストループ効果と呼ばれている (Stroop 1935)。

Botvinick らは、ACC のモニタリング機能に注目した認知制御の神経モデルを提案した (図 3.5) (Botvinick, Braver et al. 2001; Botvinick, Cohen et al. 2004; Shenhav, Botvinick et al. 2013)。ここで、このモデルがバイアス付きベイズ推定の神経モデルの一例として解釈できることを示す (Rao 2005)。刺激が提示されたときに、単語をコードしている神経細胞集団と色をコードしている神経細胞集団が活動する。ここで、これらの神経細胞集団が提示された刺激の単語と色の対数尤度をコードしていると見なすことにする。これらの層からの入力を受ける次の層の活動は課題における適切な反応が「みどり」、「あか」などであることの事後確率をコードしていると見なすことができる。

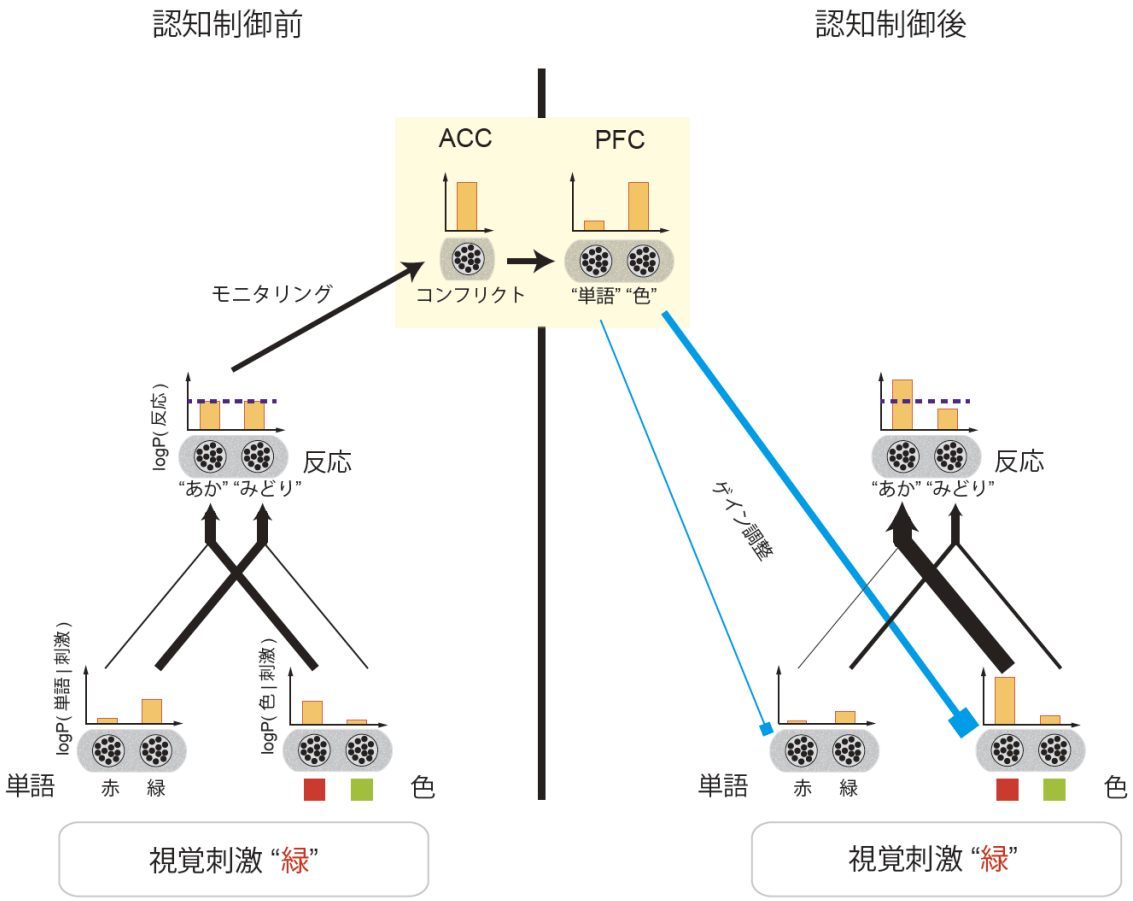

図 3.5. バイアス付きベイズ推定と認知制御。ストループ課題における認知制御の神経モデル (Botvinick et al., 2001)。このモデルはバイアス付きベイズ推定の神経モデルと同等のものと見なすことができる。

Botvinick らのモデルによると、日常生活において色の名前を言うよりも、文字の内容を読む方の頻度が高いため、色をコードしている細胞群と反応をコードしている層の結合強度よりも、単語をコードしている層と反応をコードしている層の結合強度の方が強いとされる (Botvinick, Braver et al. 2001)。

赤色で書かれた緑という文字のように色と文字が異なる刺激の場合、反応をコードしている層の「みどり」に対応する神経細胞だけでなく「あか」に対応する細胞も活動することになる。そして、文字と色が一致している場合に比べて反応をコードしている層の二つの発火率は近いものになる。異なった反応に対応する神経細胞の発火率が近いときにコンフリクトがあると見なせて、コンフリクトをモニターしている ACC が活動することになる。ACC は PFC に認知制御シグナルを送り、PFC は文字と色をコードしている層の発火率のゲインを調整するトップダウンシグナルを送る。このゲイン調整は指数バイアスの強さ (β_{word}, β_{color}) を調整していることに対応している (図 3.5)。単語読み上げ課題においては単語をコードしている神経細胞集団の発火率のゲインを上げて、色の名前課題では色をコードしている神経細胞集団の発火率のゲインを上げることで適切な反応をとれるようにしている。

- ワーキングメモリへの適用

認知制御は課題に関係ある情報を保持して、課題に無関係な情報を抑制するような

目的志向的で柔軟な行動とる際に必要とされる (Baddeley 1986)。

ここで、ワーキングメモリとバイアス付きベイズ推定の関係を議論する。トップダウンな注意によってワーキングメモリを制御することは、神経積分器のゲインすなわちバイアスの強さを調整していることに対応している。

ワーキングメモリにおいては数秒以上の間、神経活動が続く必要がある (Miller, Erickson et al. 1996; Miyake and Shah 1999)。トップダウンな注意に基づく長い間続く神経活動は PFC 内部の再帰的な結合によって保持されていると考えられている (Curtis and D'Esposito 2003; Curtis and Lee 2010)。

PFC からのトップダウンな注意のシグナルはワーキングメモリを向上させることが知られている (Desimone and Duncan 1995; Miller and Cohen 2001; Noudoost, Chang et al. 2010; Gazzaley and Nobre 2012)。このことは、ワーキングメモリにおいてトップダウンシグナルがゲインを調整するように働いていることを示唆している。この考えによると、トップダウンシグナルは α か β を調整するのに貢献していると思われる (Reynolds and Heeger ; Norman and Shallice 1986; Desimone and Duncan 1995; McAdams and Maunsell 1999; Botvinick, Braver et al. 2001; Miller and Cohen 2001; Egner and Hirsch 2005; Maunsell and Treue 2006; Cole, Reynolds et al. 2013; Cole, Repovs et al. 2014)。そのため、これは、神経積分器のバランスを変えていると見なすことができる (Gazzaley and Nobre, 2012; Roe et al., 2001)。

- トップダウン注意への適用

トップダウン注意は感覚皮質の神経活動のチューニングカーブのゲインを上げることが知られている (McAdams and Maunsell 1999; Treue and Trujillo 1999; Maunsell and Treue 2006)。チューニングカーブのゲイン上昇はノイズを減らすことだと見なすことも (Dayan and Zemel 1999)、バイアスの強さ (β) を変えているのだともみなすことができる。

チューニングカーブは課題に無関係な刺激の特性に関しても完全にフラットになるわけではなく、トップダウン注意を必要とする意思決定をする際に、無関係な情報の影響を受けてしまうことが繰り返し示されてきた (Stroop 1935; Eriksen and Eriksen 1974)。これは通常のベイズ推定では説明することができないが、指数バイアスを含んだベイズ推定において、課題に無関係な特性に関する β が 0 より大きいとすることによって説明することが可能である。これは、Biased Competition Theory のベイズ版と見なすこともできる (Desimone and Duncan 1995; Desimone 1998)。

どのようにバイアスの強さが決まっているかは、とても重要な問題であり (Shenhav, Botvinick et al. 2013)、注意を事前分布として扱っている研究も存在する (Angela and Dayan 2004; Chikkerur, Serre et al. 2010) がこれらについて扱うのは本章の範囲を超えている。

神経修飾／伝達物質及びゲイン調整と精神医学の関係性

精神医学への応用もまた重要である (Friston, Kilner et al. 2006; Friston 2009; Friston 2010; Mante, Sussillo et al. 2013; Miller and Buschman 2013)。なぜなら、ある種の精神疾患は認知制御の無効化に伴う PFC とその他の脳部位の結合が変化によって生じると考えられているからである (Cole, Repovs et al. 2014; Stephan, Iglesias et al. 2015)。

認知制御によるゲイン調整のメカニズムとして有望な仮説はノルエピネフリン、アセチルコリン、グルタミン酸、γ-アミノ酪酸 (GABA) などの神経修飾／伝達物質による調整である (Friston, Kilner et al. 2006; Friston 2009; Friston 2010)。

- ノルエピネフリン

青斑核にあるノルエピネフリン作動性の神経細胞は ACC で表現されている課題の価値の情報を受け取り、判断や意思決定に関係する細胞のゲインを上昇させている (Servan-Schreiber, Printz et al. 1990; Aston-Jones and Cohen 2005; Yu and Dayan 2005)。このゲイン調整はバイアス付きベイズ推定の指数バイアス の強さ (α と β) を変えていると見なすことができて、よりノルエピネフリン濃度が濃いときほど分布がシャープになっている (図 3.6)。

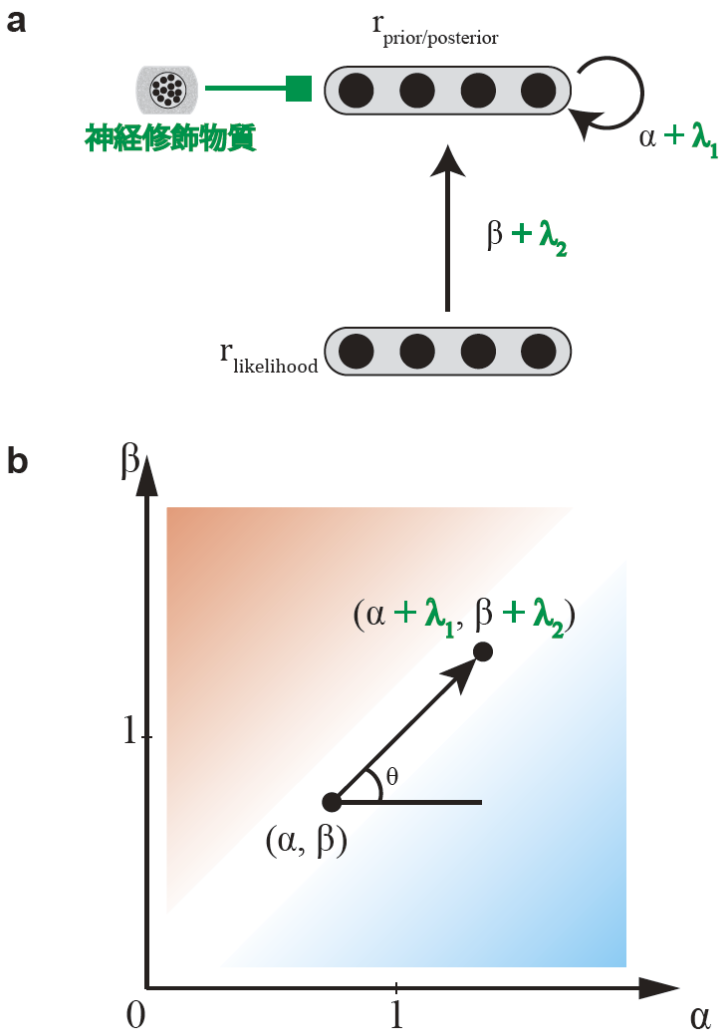

図 3.6 神経修飾／伝達物質の効果。(a) 神経修飾／伝達物質によるバイアス付きベイズ推定の神経モデルへの効果。最初の層は尤度をコードしており、二番目の層は事前／事後分布をコードしている。再帰的結合の強さは $\alpha + \lambda_1$ で、フィードフォワード結合の強さは $\beta + \lambda_2$。λ_1 と λ_2 が神経修飾／伝達物質の効果の分である。(b) バイアス平面上でみる神経修飾／伝達物質の効果。神経修飾／伝達物質はバイアスの強さを (α, β) から $(\alpha + \lambda_1, \beta + \lambda_2)$ に変える。ノルエピネフリン濃度が増えると $\lambda_1, \lambda_2 \geq 0$ $(0° \leq \theta \leq 90°)$ の変化が見込まれる。また、アセチルコリンは $\lambda_1 \leq 0, \lambda_2 \geq 0$ $(90° \leq \theta \leq 180°)$、E/I バランスにおいて E が強いときは $\lambda_1, \lambda_2 \geq 0$ $(0° \leq \theta \leq 90°)$、I が強いときは $\lambda_1, \lambda_2 \leq 0$ $(180° \leq \theta \leq 270°)$ という変化が見込まれる。

注意欠陥・多動性障害 (Attention deficit/hyperactivity disorder, ADHD) の患者はノルエピネフリンの機能が低下していることが分かっている。彼らは試行ごとに異常に探索行動をとり、反応時間の分散も大きい (Frank, Santamaria et al. 2007; Hauser, Fiore et al. 2016) が、これらの行動の発見はノルエピネフリンの濃度上昇によって分布がシャープになり、反応時間の分散も小さくなることと一貫している (Frank, Santamaria et al. 2007)。

- アセチルコリン

前脳基底部の細胞から放出されるアセチルコリンは二通りの方法で神経回路のゲインを調整する (Hasselmo and McGaughy 2004; Hasselmo 2006)。アセチルコリンの濃度上昇はフィードフォワードの結合の影響を強くする一方で、記憶を保持するための再帰的結合の影響を弱くする。そのため、アセチルコリンの濃度が高いと新しい情報のコーディングが促進され、アセチルコリンの濃度が低いと既にコードされている情報を保持し続ける。このゲイン調整はバイアス平面の *ML-like – robust* 軸 (β – α 軸) に対応している (図 3.6)。アセチルコリン濃度が高いと *ML-like* な推定になり、濃度が低いと *robust* な推定になる。

アルツハイマー病の患者はアセチルコリンの機能不全に伴う、ワーキングメモリの機能障害を引き起こす (Baddeley, Bressi et al. 1991)。これは新しい入力の影響を過小評

価し (β が小)、神経積分器のバランスが取れていないこと (α > 1) に対応している。

- 興奮性結合／抑制性結合バランス (E/I バランス)

グルタミン酸と GABA はそれぞれ興奮性と抑制性の神経伝達物質の主なものである (Kandel, Schwartz et al. 2000)。これら二つの神経伝達物質のバランス (すなわち E/I バランス) は神経回路のゲインを決定する。このゲインはバイアス付きベイズ推定のバイアスの強さに対応している (図 3.6)。

統合失調症には E/I バランスの不具合によって生じる幻覚と妄想という二つの主な陽性症状がある (Wang 2001; Kehrer, Maziashvili et al. 2008; Murray, Anticevic et al. 2014)。事前知識と外部からの入力の両方に関するゲインが正しいとき、これらの症状は起きないということが提案されてきた。Predictive coding 仮説に基づき、Corlett et al. (2009) は、妄想は事前知識に関するゲインが外部性入力のそれを上回るときにおこり、幻覚は外部性入力のゲインが事前知識のそれを上回るときにおこることを提案した (Corlett, Frith et al. 2009; Fletcher and Frith 2009; Teufel, Subramaniam et al. 2015)。事前知識と外部性入力のゲインの大きさは E/I バランスに依存しているので、幻覚と妄想は異常な E/I バランスによって説明される。さらに、Jardri や Deneve らは、事前知識と外部性入力の循環推論が統合失調症の陰性症状と陽性症状にそれぞれ対応していると述べている (Jardri and Deneve 2013; Jardri, Duverne et al. 2017)。

自閉症は社会的コミュニケーションや社会的インタラクションさらには知覚の異常といった症状のある、遺伝性の疾患である (Pellicano and Burr 2012)。ある種の自閉症は E/I バランスの異常によって引き起こされると考えられている (Rubenstein and Merzenich 2003; Nelson and Valakh 2015)。E/I バランスはバイアス付きベイズ推定のバイアスの強さを決めるので、自閉症患者は不適切な推論をおこなっているものと考えられる (図 3.3)。ここからは自閉症患者における知覚異常に焦点を当てて議論を進める。

我々の知覚が Kanizsa の三角形 (Kanizsa 1955) や Shepard のテーブル (Shepard 1990) のような視覚イリュージョンを経験することからわかるように、脳はあいまいな知覚と強い事前知識を組み合わせることで解釈をおこなっている (Pellicano and Burr 2012)。一方で、自閉症の患者はあまりイリュージョンを経験しないことから、事前分布がフラットになっていることが示唆される (Pellicano and Burr 2012)。しかしながら、Brock (2012) によると、同じ結果は尤度がシャープになることによっても得られる (Brock 2012)。この自閉症患者の推論に関するこの二つの立場は、図 3.2 の ML-like な推論を引き起こす二つのバイアスに対応しているため、バイアス平面を用いた分析が有用であると考えられる。

5. 結論

本章では、指数バイアス付きベイズ推定について議論した。バイアス付きベイズ推

定のバイアス平面には、異なったパラメータ推定法や基準率の無視、代表性バイアス、保守性バイアス、アンカリングと調整、primacy/recency 効果、選好逆転といった心理バイアスを位置づけることができた。さらに、バイアスの強さが認知制御によって決められていることを示し、その機能不全として ADHD、アルツハイマー病、統合失調症、自閉症と言った精神疾患の症状が説明できることを明らかにした。期待効用理論に、このようなバイアスを付けた推論を組み合わせることでより多くの人の行動を説明できる。

第 4 章 まとめ

第 1 章「序論」では、意思決定とは何かについて説明し、人やその他の動物がどのように意思決定しているのかについて、工学、統計学、神経科学、経営学、心理学などさまざまな分野で研究がなされてきたことについて述べた。また、合理的意思決定の理論において最もポピュラーなものである、主観的価値を最大化するように行動を決定する効用最大化アプローチについて説明した。また、行動の結果が確率的に決まる環境における意思決定において用いられる、期待効用最大化の基準について述べ、期待効用を知るのに必要な確率をベイズ推定によって計算できることについても説明した。また、この方法で期待効用理論と確率に関するベイズ推定とを組み合わせることで、さまざまな分野で別々におこなわれてきた意思決定研究は統合的に再解釈することができるため、近年では一つの研究分野と見なされるようになってきていることについて触れた。

第 2 章「社会心理学への適用」では、人々の望ましくない習慣や中毒的な行動を変える、行動変容の問題を取り扱った。行動変容は、さまざまな分野で長い間研究されており、多くの既存の行動変容のモデルでは、行動の意図を生じさせるための重要な要因として、態度、規範、自己効力感の 3 つが用いられていることについて述べて、そのような行動変容モデルの典型例として計画的行動理論について説明した。一方、行動変容モデルの精度を向上させるために、既存の行動変容モデルと行動経済学の成

果を組み合わせる試みがおこなわれつつあるが、この試みは計画的行動理論などの既存の行動変容モデルが、多くの行動経済学的モデルの基盤となっている期待効用理論と整合的ではないため難しいものになっていることについても述べた。この問題を解決するために、計画的行動理論と期待効用理論の構成要素の対応関係を明らかにした上で、期待効用理論の自然な拡張として、決定理論的な行動変容モデルを提案した。さらに、決定理論的な行動変容モデルは、通常の期待効用理論に主観的規範と自己効力感の概念を付け加えたものであることを述べ、計画的行動理論に対する優位性についても議論した。

第３章「ベイズ推定のバイアスのモデル化」では、期待効用理論と確率に関するベイズ推論とを組み合わせたモデルから人間の意思決定が逸脱する場合について検討した。そのために、ベイズ推定からの逸脱を定量的に記述するための指数型バイアス付きベイズ推論モデルについて議論した。ベイズ推定の二つの構成要素である尤度と事前確率に関する指数バイアスをそれぞれ考え、二種類のバイアスの強さに基づく「バイアス平面」を考えることができることを示すとともに、バイアス平面上に最尤推定、ベイズ推定、MAP 推定を位置づけることができることを明らかにした。また、基準率の無視や代表性バイアス、保守性バイアス、アンカリングと調整といった確率推論のバイアスもバイアス平面上に位置づけられることを示した。さらに、既存のベイズ推論の神経モデルである確率的ポピュレーションコーディングモデルにおけるシナプス

入力のゲインを変更することで、指数型バイアスを簡単に実現できることを明らかにした。これにより、DFT や LCA などの知覚的意思決定を説明する神経モデルをベイズ推論の立場から解釈できることを示した。さらに、認知制御が指数バイアスの強さを調整するメカニズムとして機能している可能性について触れ、これが神経修飾／伝達物質の調整によって実現されていることについて述べた。最後に、バイアス付きベイズ推定と ADHD やアルツハイマー病、統合失調症、自閉症などのさまざまな疾患との関連についても考察した。

以上により、意思決定の記述的なモデルである、バイアスを含むように拡張された期待効用理論を用いることで、精神疾患なども含んだより多くの人や動物の行動を表現できることが明らかとなった。ある種のバイアスには見方によっては適応的な意味があると見なせるとされているが、適応的な意味があるとされる状況においてほどバイアスがロバストであるといったような関係が見られるのかどうか、このような関係見られた場合は、有用性とロバストさの関係と精神疾患はどのように関連しているのか、といった問題についての今後の研究がまたれる。本研究がそのような進展の一助となれば幸いである。

謝辞

本論文の研究及び執筆にあたり、ご指導ご鞭撻を賜りました東京工業大学小池康晴

教授及び玉川大学脳科学研究所松元健二教授に心より御礼申し上げます。

参考文献

Agostinelli, C. and L. Greco (2012). Weighted likelihood in Bayesian inference. 46TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY.

Agostinelli, C. and L. Greco (2013). "A weighted strategy to handle likelihood uncertainty in Bayesian inference." Computational Statistics **28**(1): 319-339.

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. Action Control: From Cognition to Behavior. J. Kuhl and J. Beckmann. Berlin, Heidelberg, Springer Berlin Heidelberg: 11-39.

Ajzen, I. (1991). "THE THEORY OF PLANNED BEHAVIOR." Organizational Behavior and Human Decision Processes **50**(2): 179-211.

Angela, J. Y. and P. Dayan (2004). Inference, attention, and decision in a Bayesian neural architecture. Advances in neural information processing systems.

Aston-Jones, G. and J. D. Cohen (2005). "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance." Annu Rev Neurosci **28**: 403-450.

Baddeley, A. (1986). "Working Memory."

Baddeley, A. D., S. Bressi, et al. (1991). "The decline of working memory in Alzheimer's disease. A longitudinal study." Brain **114 (Pt 6)**: 2521-2542.

Bandura, A. (1982). "Self-efficacy mechanism in human agency." American Psychologist **37**(2): 122-147.

Barlow, H. B. (1969). "Pattern recognition and the responses of sensory neurons." Ann N Y Acad Sci **156**(2): 872-881.

Beck, Jeffrey M., Wei J. Ma, et al. (2012). "Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability." Neuron **74**(1): 30-39.

Bitzer, S., H. Park, et al. (2014). "Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model." Front Hum Neurosci **8**: 102.

Bogacz, R., E. Brown, et al. (2006). "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." Psychol Rev **113**(4): 700-765.

Botvinick, M. M., T. S. Braver, et al. (2001). "Conflict monitoring and cognitive control." Psychol Rev **108**(3): 624-652.

Botvinick, M. M., J. D. Cohen, et al. (2004). "Conflict monitoring and anterior cingulate cortex: an update." Trends in Cognitive Sciences **8**(12): 539-546.

Bratman, M. (1987). Intention, Plans, and Practical Reason, Center for the Study of Language and Information.

Brock, J. (2012). "Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr." Trends in Cognitive Sciences **16**(12): 573-574.

Brunton, B. W., M. M. Botvinick, et al. (2013). "Rats and Humans Can Optimally Accumulate Evidence for Decision-Making." Science **340**(6128): 95-98.

Busemeyer, J. R. and J. T. Townsend (1993). "Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment." Psychological review **100**(3): 432.

Chandrasekaran, C. (2017). "Computational principles and models of multisensory integration." Current opinion in neurobiology **43**: 25-34.

Chikkerur, S., T. Serre, et al. (2010). "What and where: A Bayesian inference theory of attention." Vision Research **50**(22): 2233-2247.

Churchland, A. K. and J. Ditterich (2012). "New advances in understanding decisions among multiple alternatives." Current opinion in neurobiology **22**(6): 920-926.

Cole, M. W., G. Repovs, et al. (2014). "The frontoparietal control system: a central role in mental health." Neuroscientist **20**(6): 652-664.

Cole, M. W., J. R. Reynolds, et al. (2013). "Multi-task connectivity reveals flexible hubs for adaptive task control." Nat Neurosci **16**(9): 1348-1355.

Corlett, P. R., C. D. Frith, et al. (2009). "From drugs to deprivation: a Bayesian framework

for understanding models of psychosis." Psychopharmacology **206**(4): 515-530.

Crockett, M. J. (2013). "Models of morality." Trends in Cognitive Sciences **17**(8): 363-366.

Curtis, C. E. and M. D'Esposito (2003). "Persistent activity in the prefrontal cortex during working memory." Trends Cogn Sci **7**(9): 415-423.

Curtis, C. E. and D. Lee (2010). "Beyond working memory: the role of persistent activity in decision making." Trends Cogn Sci **14**(5): 216-222.

Cushman, F. (2008). "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment." Cognition **108**(2): 353-380.

Dayan, P. and L. F. Abbott (2001). Theoretical neuroscience, Cambridge, MA: MIT Press.

Dayan, P. and R. S. Zemel (1999) "Statistical models and sensory attention." IET Conference Proceedings, 1017-1022.

Desimone, R. (1998). "Visual attention mediated by biased competition in extrastriate visual cortex." Philosophical Transactions of the Royal Society B: Biological Sciences **353**(1373): 1245-1255.

Desimone, R. and J. Duncan (1995). "Neural Mechanisms of Selective Visual Attention." Annual review of neuroscience **18**(1): 193-222.

Edwards, W. (1954). "The theory of decision making." Psychol Bull **51**(4): 380-417.

Egner, T. and J. Hirsch (2005). "Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information." Nat Neurosci **8**(12): 1784-1790.

Eriksen, B. A. and C. W. Eriksen (1974). "Effects of noise letters upon the identification of a target letter in a nonsearch task." Perception & psychophysics **16**(1): 143-149.

Ernst, M. O. and M. S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion." Nature **415**(6870): 429-433.

Ernst, M. O. and H. H. Bulthoff (2004). "Merging the senses into a robust percept." Trends Cogn Sci **8**(4): 162-169.

Fehr, E. and I. Krajbich (2014). Social Preferences and the Brain. Neuroeconomics (Second Edition). P. W. Glimcher and E. Fehr. San Diego, Academic Press: 193-218.

Fishbein, M. and I. Ajzen (2010). Predicting and changing behavior: The reasoned action approach, Taylor & Francis.

Fletcher, P. C. and C. D. Frith (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia." Nat Rev Neurosci **10**(1): 48-58.

Frank, M. J., A. Santamaria, et al. (2007). "Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder." Neuropsychopharmacology **32**(7): 1583-1599.

Friston, K. (2009). "The free-energy principle: a rough guide to the brain?" Trends Cogn Sci **13**(7): 293-301.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" Nat Rev Neurosci **11**(2): 127-138.

Friston, K., J. Kilner, et al. (2006). "A free energy principle for the brain." J Physiol Paris **100**(1-3): 70-87.

Friston, K., P. Schwartenbeck, et al. (2013). "The anatomy of choice: active inference and agency." Front Hum Neurosci **7**: 598.

Gazzaley, A. and A. C. Nobre (2012). "Top-down modulation: bridging selective attention and working memory." Trends Cogn Sci **16**(2): 129-135.

Gazzaniga, M. S. (2009). The cognitive neurosciences, MIT press.

Gigerenzer, G. and D. G. Goldstein (1996). "Reasoning the fast and frugal way: models of bounded rationality." Psychological review **103**(4): 650.

Gilboa, I. (2010). Making better decisions: Decision theory in practice, John Wiley & Sons.

Gilboa, I. (2012). Rational choice.

Glimcher, P. W. (2003). "The neurobiology of visual-saccadic decision making." Annu Rev Neurosci **26**: 133-179.

Gold, J. I. and M. N. Shadlen (2001). "Neural computations that underlie decisions about sensory stimuli." Trends in Cognitive Sciences **5**(1): 10-16.

Goldman, M. S., A. Compte, et al. (2009). Neural Integrator Models. Encyclopedia of

Neuroscience. L. R. Squire. Oxford, Academic Press: 165-178.

Grether, D. M. (1980). "Bayes rule as a descriptive model: The representativeness heuristic." The Quarterly Journal of Economics: 537-557.

Hardeman, W., M. Johnston, et al. (2002). "Application of the Theory of Planned Behaviour in Behaviour Change Interventions: A Systematic Review." Psychology & Health **17**(2): 123-158.

Hasselmo, M. E. (2006). "The Role of Acetylcholine in Learning and Memory." Current opinion in neurobiology **16**(6): 710-715.

Hasselmo, M. E. and J. McGaughy (2004). "High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation." Prog Brain Res **145**: 207-231.

Hauser, T. U., V. G. Fiore, et al. (2016). "Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis." Trends Neurosci **39**(2): 63-73.

Jardri, R. and S. Deneve (2013). "Circular inferences in schizophrenia." Brain **136**(Pt 11): 3227-3241.

Jardri, R., S. Duverne, et al. (2017). "Experimental evidence for circular inference in schizophrenia." Nat Commun **8**: 14218.

Jazayeri, M. and J. A. Movshon (2006). "Optimal representation of sensory information by neural populations." Nature Neuroscience **9**(5): 690-696.

Körding, K. P. and D. M. Wolpert (2004). "Bayesian integration in sensorimotor learning." Nature **427**(6971): 244-247.

Kahneman, D., P. Slovic, et al. (1982). Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." Econometrica **47**(2): 263-291.

Kandel, E. R., J. H. Schwartz, et al. (2000). Principles of neural science, McGraw-Hill New York.

Kanizsa, G. (1955). "Margini quasi-percettivi in campi con stimolazione omogenea." Rivista di psicologia **49**(1): 7-30.

Kehrer, C., N. Maziashvili, et al. (2008). "Altered Excitatory-Inhibitory Balance in the NMDA-Hypofunction Model of Schizophrenia." Frontiers in Molecular Neuroscience **1**: 6.

Kiani, R., T. D. Hanks, et al. (2008). "Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment." The Journal of Neuroscience **28**(12): 3017-3029.

Kira, S., T. Yang, et al. (2015). "A neural implementation of Wald's sequential probability ratio test." Neuron **85**(4): 861-873.

Krajbich, I., C. Armel, et al. (2010). "Visual fixations and the computation and comparison of value in simple choice." Nat Neurosci **13**(10): 1292-1298.

Kulhavý, R. and M. B. Zarrop (1993). "On a general concept of forgetting." International Journal of Control **58**(4): 905-924.

Luce, R. D. (1959). Individual choice behavior. Oxford, England, John Wiley.

Ma, W. J., J. M. Beck, et al. (2006). "Bayesian inference with probabilistic population codes." Nature Neuroscience **9**(11): 1432-1438.

Ma, W. J. and M. Jazayeri (2014). "Neural coding of uncertainty and probability." Annual review of neuroscience **37**: 205-220.

Mante, V., D. Sussillo, et al. (2013). "Context-dependent computation by recurrent dynamics in prefrontal cortex." Nature **503**(7474): 78-84.

Matsumori, K., K. Iijima, et al. (2019). "A Decision-Theoretic Model of Behavior Change." Frontiers in Psychology **10**(1042).

Matsumori, K., Y. Koike, et al. (2018). "A Biased Bayesian Inference for Decision-Making and Cognitive Control." Frontiers in Neuroscience **12**(734).

Matsumoto, K. (2004). "Conflict and cognitive control." Science **303**(5660): 969-970.

Maunsell, J. H. and S. Treue (2006). "Feature-based attention in visual cortex." Trends

Neurosci **29**(6): 317-322.

Mazurek, M. E., J. D. Roitman, et al. (2003). "A Role for Neural Integrators in Perceptual Decision Making." Cerebral Cortex **13**(11): 1257-1269.

McAdams, C. J. and J. H. Maunsell (1999). "Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4." J Neurosci **19**(1): 431-441.

Miller, E. K. and T. J. Buschman (2013). "Cortical Circuits for the Control of Attention." Current opinion in neurobiology **23**(2): 216-222.

Miller, E. K. and J. D. Cohen (2001). "An integrative theory of prefrontal cortex function." Annu Rev Neurosci **24**: 167-202.

Miller, E. K., C. A. Erickson, et al. (1996). "Neural mechanisms of visual working memory in prefrontal cortex of the macaque." J Neurosci **16**(16): 5154-5167.

Miyake, A. and P. Shah (1999). Models of working memory: Mechanisms of active maintenance and executive control, Cambridge University Press.

Mulder, M. J., E.-J. Wagenmakers, et al. (2012). "Bias in the brain: a diffusion model analysis of prior probability and potential payoff." The Journal of Neuroscience **32**(7): 2335-2343.

Murray, J. D., A. Anticevic, et al. (2014). "Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model." Cereb Cortex **24**(4): 859-872.

Nassar, M. R., R. C. Wilson, et al. (2010). "An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment." The Journal of Neuroscience **30**(37): 12366-12378.

Nelson, Sacha B. and V. Valakh (2015). "Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders." Neuron **87**(4): 684-698.

Neyman, J. and E. S. Pearson (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **231**(694-706): 289-337.

Norman, D. A. and T. Shallice (1986). Attention to action, Springer.

Noudoost, B., M. H. Chang, et al. (2010). "Top-down control of visual attention." Current opinion in neurobiology **20**(2): 183-190.

Payzan-LeNestour, E. and P. Bossaerts (2011). "Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings." PLoS Comput Biol **7**(1): e1001048.

Payzan-LeNestour, É. and P. Bossaerts (2012). "Do not Bet on the Unknown Versus Try to Find Out More: Estimation Uncertainty and "Unexpected Uncertainty" Both Modulate Exploration." Frontiers in Neuroscience **6**: 150.

Payzan-LeNestour, E., S. Dunne, et al. (2013). "The neural representation of unexpected uncertainty during value-based decision making." Neuron **79**(1): 191-201.

Pellicano, E. and D. Burr (2012). "Response to Brock: noise and autism." Trends in Cognitive Sciences **16**(12): 574-575.

Pellicano, E. and D. Burr (2012). "When the world becomes "too real": a Bayesian explanation of autistic perception." Trends in Cognitive Sciences **16**(10): 504-510.

Peterka, V. (1981). "Bayesian approach to system identification." Trends and Progress in System identification **1**: 239-304.

Pouget, A., J. M. Beck, et al. (2013). "Probabilistic brains: knowns and unknowns." Nature Neuroscience **16**(9): 1170-1178.

Rangel, A., C. Camerer, et al. (2008). "A framework for studying the neurobiology of value-based decision making." Nat Rev Neurosci **9**(7): 545-556.

Rao, R. P. N. (2005). "Bayesian inference and attentional modulation in the visual cortex." NeuroReport **16**(16): 1843-1848.

Ratcliff, R. (1978). "A theory of memory retrieval." Psychological review **85**(2): 59.

Redish, A. D. (2004). "Addiction as a computational process gone awry." Science **306**(5703): 1944-1947.

Reynolds, J. H. and D. J. Heeger "The Normalization Model of Attention." Neuron **61**(2):

168-185.

Roberto, C. A. and I. Kawachi (2015). Behavioral economics and public health, Oxford University Press.

Roe, R. M., J. R. Busemeyer, et al. (2001). "Multialternative decision field theory: A dynamic connectionst model of decision making." Psychological review **108**(2): 370.

Rorie, A. E., J. Gao, et al. (2010). "Integration of Sensory and Reward Information during Perceptual Decision-Making in Lateral Intraparietal Cortex (LIP) of the Macaque Monkey." PLoS ONE **5**(2): e9308.

Rubenstein, J. L. and M. M. Merzenich (2003). "Model of autism: increased ratio of excitation/inhibition in key neural systems." Genes Brain Behav **2**(5): 255-267.

Schoemaker, P. (1982). "The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations." Journal of Economic Literature **20**(2): 529-563.

Servan-Schreiber, D., H. Printz, et al. (1990). "A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior." Science **249**(4971): 892-895.

Shadlen, M. N. and W. T. Newsome (2001). "Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey." J Neurophysiol **86**(4): 1916-1936.

Sheeran, P., A. Maki, et al. (2016). "The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: A meta-analysis." Health Psychology **35**(11): 1178-1188.

Shenhav, A., M. M. Botvinick, et al. (2013). "The expected value of control: an integrative theory of anterior cingulate cortex function." Neuron **79**(2): 217-240.

Shepard, R. N. (1990). Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art, WH Freeman/Times Books/Henry Holt & Co.

Sniehotta, F. F., J. Presseau, et al. (2014). "Time to retire the theory of planned behaviour." Health Psychology Review **8**(1): 1-7.

Soltani, A. and X.-J. Wang (2010). "Synaptic computation underlying probabilistic

inference." Nature Neuroscience **13**(1): 112-119.

Sonnenberg, F. A. and J. R. Beck (1993). "Markov Models in Medical Decision Making:A Practical Guide." Medical Decision Making **13**(4): 322-338.

Stephan, K. E., S. Iglesias, et al. (2015). "Translational Perspectives for Computational Neuroimaging." Neuron **87**(4): 716-732.

Stone, J. V. (2013). Bayes' rule: a tutorial introduction to Bayesian analysis, JV Stone.

Stone, M. (1960). "Models for choice-reaction time." Psychometrika **25**(3): 251-260.

Story, G. W., I. Vlaev, et al. (2014). "Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective." Frontiers in Behavioral Neuroscience **8**: 76.

Stroop, J. R. (1935). "Studies of interference in serial verbal reactions." Journal of experimental psychology **18**(6): 643.

Summerfield, C. and E. Koechlin (2010). "Economic Value Biases Uncertain Perceptual Choices in the Parietal and Prefrontal Cortices." Frontiers in Human Neuroscience **4**: 208.

Summerfield, C. and K. Tsetsos (2012). "Building bridges between perceptual and economic decision-making: neural and computational mechanisms." Frontiers in Neuroscience **6**.

Sutton, R. S. and A. G. Barto (1998). Reinforcement learning: An introduction.

Teufel, C., N. Subramaniam, et al. (2015). "Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals." Proceedings of the National Academy of Sciences **112**(43): 13401-13406.

Treue, S. and J. C. M. Trujillo (1999). "Feature-based attention influences motion processing gain in macaque visual cortex." Nature **399**(6736): 575-579.

Tsetsos, K., J. Gao, et al. (2012). "Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model." Front Neurosci **6**: 79.

Tversky, A. and D. Kahneman (1974). "Judgment under uncertainty: Heuristics and biases." Science **185**(4157): 1124-1131.

Usher, M. and J. L. McClelland (2001). "The time course of perceptual choice: the leaky, competing accumulator model." Psychological review **108**(3): 550.

Usher, M. and J. L. McClelland (2004). "Loss aversion and inhibition in dynamical models of multialternative choice." Psychological review **111**(3): 757.

Van Lange, P., A. Kruglanski, et al. (2011). Handbook of Theories of Social Psychology SAGE Publications Ltd.

von Neumann, J. and O. Morgenstern (1947). Theory of Games and Economic Behavior, Princeton university press.

Wald, A. (1973). Sequential analysis, Courier Corporation.

Wald, A. and J. Wolfowitz (1948). "Optimum character of the sequential probability ratio test." The Annals of Mathematical Statistics: 326-339.

Wang, X.-J. (2001). "Synaptic reverberation underlying mnemonic persistent activity." Trends in Neurosciences **24**(8): 455-463.

Wang, X.-J. (2002). "Probabilistic Decision Making by Slow Reverberation in Cortical Circuits." Neuron **36**(5): 955-968.

Wolpert, D. M., Z. Ghahramani, et al. (1995). "An internal model for sensorimotor integration." Science-AAAS-Weekly Paper Edition **269**(5232): 1880-1882.

Wong, K. F. and X. J. Wang (2006). "A recurrent network mechanism of time integration in perceptual decisions." J Neurosci **26**(4): 1314-1328.

Yang, T. and M. N. Shadlen (2007). "Probabilistic reasoning by neurons." Nature **447**(7148): 1075-1080.

Yu, A. J. and P. Dayan (2005). "Uncertainty, neuromodulation, and attention." Neuron **46**(4): 681-692.