

## ビッグデータ倫理<sup>1</sup>

—生物医学の文脈における現在のそして予見される問題について—

ブレント・ミッテルシュッタ<sup>2</sup>、ルチアーノ・フロリディ<sup>3</sup>

### 出典：

Brent Mittelstadt, Luciano Floridi, “The Ethics of Data: Current and Foreseeable Issues in Biomedical Contexts”, in *Science and Engineering Ethics*, 2015.

### キーワード：

- ・ 倫理(Ethics)
- ・ ビッグデータ(Big data)
- ・ 生命倫理(Bioethics)
- ・ 情報倫理(Information ethics)
- ・ 医療倫理(Medical ethics)
- ・ 倫理的展望(Ethical foresight)

### 凡例：

本文中の“ ”は「 」で、イタリックなどの強調は傍点にて示す。また、( )は原語表記などのために用いられている。なお、補足的な説明は脚注をつけて行うこととする。

### 論文の目的

はじめに、紹介者により本論文の目的について述べておく。昨今、ビッグデータを収集、蓄積し、分析する能力が飛躍的に向上している。しかし、その進歩から期待される利益ばかりに注目が集まり、社会的、倫理的影響に対する理解が遅れている。本論文では、ビッグデータの倫理的問題やそのリスクについて、問題の生じやすい生物医学のビッグデータに焦点をあてつつ、先行研究を調査することで明らかにする。また、その後で、先行研究で見落とされていたり、今後重要性を増すと思われる問題を指摘する。この作業を通して、今後の

---

<sup>1</sup> Brent Mittelstadt, Luciano Floridi, “The Ethics of Data: Current and Foreseeable Issues in Biomedical Contexts”, in *Science and Engineering Ethics*, 2015.

<sup>2</sup> オックスフォード大学インターネット研究所首席研究員。専門はデータ倫理、AI、アルゴリズム、機械学習、プロファイリング、医療倫理等。

<sup>3</sup> オックスフォード大学教授

ビッグデータ実践の倫理的、また責任ある発展に寄与することを目的とする。

## アブストラクト（全訳）

データを収集し、分析する能力は指数関数的に向上している。「ビッグデータ」といわれる科学的、社会的、技術的趨勢は、受け入れられている社会的ないし倫理的規範に異議を唱えられるほど、不安定な量の情報を生み出すことを促進している。しかしビッグデータは、時に巨大なデータセットの規模のみが関係しているかのような仕方でも社会的、科学的、商業的な文脈にまたがって用いられていることもあり、依然曖昧な概念である。また、科学や技術の最先端ではよく見られるように、ビッグデータの倫理的影響についての理解は遅れている。このようなギャップを埋めるため、本稿ではビッグデータの倫理的影響に関する学術文献を体系的かつ包括的に分析し、これからの倫理的調査や規制に分水嶺(watershed)を与える。とりわけ、生物医学のビッグデータは、医療情報に固有の機密性があることから、特に注意が必要である。文献のメタ分析により、新しく、そして革新的な現象が持つ倫理的リスクに関して、すでに知られていること、ないしは仮説が立てられていることを通して、倫理学者、データ科学者、規制当局、そしてその他の利害関係者に対する指針を示すため、テーマごとに説明を与える。懸念されるテーマとして 5 つの主要な分野が特定される。すなわち、(1) インフォームドコンセント、(2) プライバシー（匿名化とデータ保護を含む）、(3) 所有権、(4) 認識論と客観性、(5) ますま規模の大きくなるデータセットを分析するために必要なリソースを持っている者とそうでない者の間に生じるビッグデータ格差、以上の 5 つである。これらのテーマを扱う中で、重大なギャップが明らかにされ、これからの研究への提案として示される。続いて、関連してはいるものの、既存の文献ではまだ広範な議論がなされていない 6 つの分野を提示する。これらは、ごく近い将来において、より徹底した綿密な調査を必要とするだろう。この 6 つとは、(6) グループレベルの倫理的危害を無視することの危険性、(7) ビッグデータ倫理の評価における認識論の重要性、(8) ますまデータであふれる信頼関係の特性の変質、(9) データ主体への潜在的な危害という観点から、「学術的」ビッグデータ実践と、「商業的」ビッグデータ実践を区別する必要性、(10) 集約されたデータセットの分析から生じた知的財産権に関する将来的問題、(11) 必要なリソースが不足している個々のデータ主体に意味のあるアクセス権を提供することの困難さ、である。これら 11 のテーマを考え合わせることで、新興のビッグデータ実践における倫理的評価とガバナンスの指針となる、完全な批判的枠組みが得られる。

## 論文の構成

- 1、調査方法の紹介
- 2、調査結果、先行研究での議論のまとめ
- 3、重要な問題の指摘

## 1、調査方法の紹介

実施時期：2014年10月

方法

- ① 6つのデータベース(Web of Science、Scopus、Global Health、Philpapers、PubMed Google Scholar)で、ビッグデータの倫理問題を扱う論文を検索した。その際、生物医学に関するものと、そうでないものを特定できるよう、区別して行った。検索に用いたキーワードについては、表1を参照されたい。

表1

Database	Search string	Returned
<b>Ethics of biomedical big data</b>		
Web of science	TOPIC: ((ethic* OR moral*) (health* OR *medic* OR bio*) 'big data')	23
Scopus	TITLE-ABS-KEY: ((ethic* OR moral*) AND (health* OR *medic* OR bio*) AND 'big data')	18
Global health	'Big data' AND (ethic* OR moral*)	145
PubMed	'Big data' AND (ethic* OR moral*)	19
<b>Ethics of big data</b>		
Web of science	TOPIC: ((ethic* OR moral*) 'big data' NOT (*medic* OR health* OR bio*))	18
Scopus	TITLE-ABS-KEY: ((ethic* OR moral*) AND 'big data' AND NOT (*medic* OR health* OR bio*))	28
Philpapers	'Big data'	19
Google scholar	'Big data' ethics OR ethical OR ethic OR moral OR morality OR morals	50 <sup>a</sup>
Philosopher's index	'Big data'	6

<sup>a</sup> 11,000 Returned, first 50 reviewed

② 検索された論文のタイトルと、要旨をもとに関連性を調査し、倫理的議論を行っているものを選出した。ただし、文献は英語論文に限定した。また査読論文の他に、白書や科学書等の出版物も含む。なお、期限に制限を設けなかった。

③ 選出された文献の分析を行う。最初に、本文中で倫理的問題や、考えが表れているフレーズを強調する。その後、それらをグループ分けし、倫理的テーマを割り当てた。これらのテーマが出揃った後で、NVivo 10 ソフトウェアを用いて二度目の体系的分析を行った。また、

用語は統一し、ビッグデータに関する利害関係者について次のように定める<sup>4</sup>。

データ主体(data subject)：データによって表される個人

データ管理者(data custodian)：データの運用、保管に責任を負う個人ないし組織

データ分析者(data analyst)：データを分析する個人ないし、組織（ただし、データを運用しているかは問わない）

## 2、調査結果

結果（文献内訳）

選出された文献は（2005～2014）68本であり、そのうち生物医学、健康、医療に関するものは36本、一般的なビッグデータの倫理問題を扱うものは32本であった。これらの文献の分析を行った結果、5つのテーマを特定した。すなわち、①インフォームドコンセント、②プライバシー、③所有権、④認識論、⑤ビッグデータ格差である。以下でそれぞれの問題、議論をまとめる。

ビッグデータとは何か

1、最も多くみられた定義は、4つのVを用いたものである。すなわち、(1)量〈Volume〉、データ量が非常に大きいこと、(2)速度、ないしは頻度〈Velocity〉、ストリーミングデータの分析速度・頻度が大きいこと、(3)多様性〈Variety〉、様々な形態のデータがあること、(4)正確性〈Veracity〉、データに不確実性があること、以上の4つである。

2、より広義には、ビッグデータは(1)大量のデータセットを分析する過程、ないしは(2)データセットそれ自体を指しうる。それゆえ、何をもって大量とするかにも、二つのアプローチがある。すなわち、テラバイト等で表されるような電子サイズそのものの大きさによって決定するアプローチと、既存のコンピューターテクノロジーによる技術では分析、処理が困難なものをビッグデータとするアプローチである。後者のアプローチに従うと、現在ビッグデータと見なされているものは、技術の発展した未来には、もはやビッグデータではなくなる。その未来においては、そのときの技術や分析のレベルに合わせて、ビッグデータが定義されることになる。ビッグデータの利用がデータを扱う技術のレベルに依存することを鑑みれば、この見方は、現在そして未来に、ビッグデータがどんな価値を持つのかを考えるうえでも有用である。

---

<sup>4</sup> 翻訳にあたり、訳語は個人情報保護委員会が行ったEU一般データ保護規則の翻訳を参考にした。

「個人データの取扱いと関連する自然人の保護に関する、及び、そのデータの自由な移転に関する、並びに、指令95/46/ECを廃止する欧州議会及び理事会の2016年4月27日の規則(EU)2016/679」。個人情報保護委員会。 <https://www.ppc.go.jp/files/pdf/gdpr-provisions-ja.pdf>（最終閲覧日2020/5/20）

## ①インフォームドコンセント

### ・同意取得の難しさ

データを集約して扱うビッグデータでは、従来の個別同意を前提とするインフォームドコンセントの枠組みでは対応できない問題が生じる。ビッグデータ研究はデータが収集されるときに、今後どのような目的で利用されるか、どのような機関と共有されるか、何が明らかになるのか不確実であることが多い。そのため、同意の取り方を変更することが提案されている。有力なものとして、段階的(tiered)同意があげられている。

### ・データの公開性

ビッグデータは利用可能性を高めるため、データの公開性(openness)が求められる。しかし公開されたデータが国境を越えて利用されることも考えられ、データ収集時にグローバルなタイプの同意が必要になる。また、匿名化して公開することも提案されるが、ビッグデータの場合は再識別化の可能性があり難しい。

### ・意図せず被験者になる危険性

ソーシャルメディアやアプリのサービスを利用するために、データが収集され、分析されることを望んでいなくても認めざるを得ない、または知らないうちに被験者となるという問題が生じることがある。

### ・同意取得の必要性を弱める主張がある

こうした状況の中、そもそも同意取得の必要性を減じようとする議論もみられる。オプトアウトの採用や、連帯に注目するアプローチ(バイオバンク)などが考えられているが、データの種類によって、なにが適切か考えられるべきである。ただし、社会的利益になるという理由から、同意権を弱めたり、放棄するというのは、個人に倫理的に問題のある重荷を負わせる可能性があることに注意する必要がある。

## ②プライバシー

### ・プライバシー理解について

プライバシーは匿名化と機密性に関連させられ、頻繁に取り上げられている。また用いられるプライバシーの理論や枠組みは、OECDの公正情報原則、またニッセンバウムの文脈的統一性(contextual integrity)が影響力を持っていた。

### ・データの長期的で広範な蓄積がリスクを高める

ビッグデータでは、インターネットを介したデータのやり取りはデフォルトで公的空間に存在する。こうした状況では、ただ公的空間にあるデータと、データ主体が意図して公的に知られたものにしたデータの区別が見過ごされてしまい、プライバシーの侵害が生じやすい。しかも、長い期間、そして広範な個人的データが蓄積され、永久にアクセスできてしまう。それゆえプライバシー侵害のリスクも高い。データ主体がこのリスクを認識し、セン

シティブなデータの 사용을禁止するなど、デジタル空間での壁の作り方を考える必要あり。

- ・匿名化について

プライバシーの保護のために、匿名化がしばしば要求される。しかし、ビッグデータは再識別が可能である可能性が高く、匿名化が困難。

- ・再識別の心配がなくても…

医学研究で見られるように、非識別化したデータの集合全体を分析するのなら、再識別はなされないという主張もある。しかし、集約されたビッグデータの分析により、集団レベルで害される可能性がある。これまでの同意に関する議論もプライバシーの議論も基本的に個人に関わるものとして考えられ、こうした集団レベルでの問題を扱うことができていない。これについては後述する必要がある。

- ・データの保護に関して

現行の個人情報保護法では全ての医学的に重要なデータを守ることができていない。データ管理者が倫理委員会等に監視されていれば、その場合でも問題は生じないが、PatientsLikeMe<sup>5</sup>のような患者主導のデータベースや、民営の健康情報サービスに適切な制約を設けることができていない。

### ③所有権

- ・データのコントロール権

データのコントロール権をデータ主体が行使できるよう、データの存在と操作を追跡、監視できるメカニズムが必要。

- ・データからの受益権

データからの受益権は、通常データ分析者が持つものと考えられているが、データ主体にも認めるべきとの主張を行う論文がひとつだけあった。重要なので後述する。

- ・アクセス能力の限界

いずれの権利を行使するにしても、データ主体がデータにアクセスする能力には限界があり、データ主体に意味のあるアクセス権を付与する必要がある。これは今後議論されるべき重要な問題であり、後述する。

### ④認識論

- ・ビッグデータの複雑さ

ビッグデータは、データそれ自体が多様、膨大で複雑であるうえに、アルゴリズムもまた複

---

<sup>5</sup> PatientsLikeMe は患者同士が体験を共有できるネットコミュニティとして 2004 年に設立されたサイトである。詳細については公式サイト [<https://www.patientslikeme.com/>] を参照のこと。

雑性を持つ。それゆえ、ビッグデータに基づく発見が妥当なのか判断することが、AI のブラックボックス問題同様難しい。

- ・客観性についての誤った見方

ビッグデータを人間の解釈を必要とせず、客観的真理を明らかにできるものを見なす傾向がある。しかし、データを扱う際には、解釈的枠組みを通した変質が必ず伴う。ビッグデータは客観的で生のデータではないということを認識する必要がある。

- ・データの質

ビッグデータは様々な仕方で収集されたデータで構成される。研究を念頭におかず、相互運用性の低いデータもある。また因果関係を特定しようと観測データを収集すると、バイアスがかかりやすい。そのため、人によるチェックによって、データの質を担保する必要がある。

- ・文脈の喪失

データが収集される時、文脈が喪失されてしまう。それゆえ、データ収集時の目的、対象、社会的状況、方法といった情報が失われてしまう。加えて、その文脈の中で持たされていた意味も失われてしまう。これにより、他の文脈では問題のないデータが、文脈を離れることで途端にセンシティブになることがありうる（例：SNS での投稿）。

これらの問題は先行研究では認識論的問題として言及されるが、倫理的にも重要な問題なので、再び取り上げる。

## ⑤ビッグデータ格差(The Big Data Divide)

- ・ビッグデータ格差とはどのような格差か

ビッグデータ格差は、主にはデータ利用のためのインフラを備えている組織と、そうでない組織やデータ主体との間に生じる格差である。ビッグデータはますます大きな組織に独占的な領域になっており、格差は深刻である。

- ・データ利用の監視、修正の必要性

また格差は、データ主体が自分のデータが作成されたり、操作されたときにそれを知ることができるか、また修正できるかという観点からも考えられる。データ主体はプロファイリングされることで、個人のアイデンティティや関心とは無関係に、一方的に付与された属性で評価されることがありうる。これが不適切な場合もあるため、組織を監視し、場合によっては、データの削除や、忘れられる権利を保障し、不均衡を是正することが必要である。組織を監視し、問題がある場合には、データを修正したり、消去することを可能にするべきである。

- ・データの局在

ビッグデータはデジタル化に適したデータ主体や、現象に偏って蓄積されている。そのため、そうでない人々との間にも格差が生じる他、見落とされる部分が生じる可能性がある。また、データ提供の機会に恵まれるか否かは経済状況とも関わり正義の問題にもなる（医学では社会的、経済的に地位のある欧米人のデータが多い）。それゆえ、分析も偏っている。ま

た、データ主体の母体を多様にしたとしても、第三世界での薬学研究のように、適切な契約を結んでいないことで、恩恵に与れないデータ主体が出てきてしまうことが予想される。

### 3、重要な問題の指摘

まず、全ての文献を通して、経験的調査の不足と、概念の深い分析が十分になされていない部分がある。また、一般的な意見や問題の概説があるものの、深い議論が行われておらず、今後の研究がまたれる。

以下では、文献調査を経て、近い将来重要となりうる問題であり、さらなる説明が必要なものを取り上げる。①集団レベルの倫理、②経験論的困難、③信頼関係、④学術的活動と商業活動、⑤知的財産権、⑥データアクセス権の順に言及する。

#### ①集団レベルの倫理

ビッグデータでは、個人がプロファイリングされ、グループ分けされることで属性を付与される。これによってその集団の構成員として選択的扱いを受ける可能性がある（パーソナライズド・プライシングや、遺伝的差別等）。これによる差別を防ぐ手立てとして、二つのアプローチが考えられる。第一に、ケア倫理によるアプローチで、特定の形態の研究や、仮説を許されないものとして禁止、無効にするというものである<sup>6</sup>。第二に、集団レベルの概念としてプライバシーを再考し、集団プライバシー権を確立するアプローチである。後者はさらなる研究が必要である<sup>7</sup>。

#### ②認識論的困難

ビッグデータでは、文脈の喪失により、データのセンシティブリティが不明確になってしまう。しかし、先行研究では認識論的問題として指摘されるにとどまり、解釈の責任を開発者でなく、アルゴリズムそのものに課している。これは倫理的問題として認められるべきであり、さらなる議論が必要である。

また、データを収集し、分析する際に質的、文脈的喪失があることは、ビッグデータに限

---

<sup>6</sup> Lewis et al. (2012)がこのアプローチを提案。

Lewis, C. M., Obregon-Tito, A., Tito, R. Y., Foster, M. W., & Spicer, P. G. (2012). The Human Microbiome Project: Lessons from human genomics. *Trends in Microbiology*, 20(1), pp. 1–4.

<sup>7</sup> ミッテルシュタットは、このアプローチを採用し“From Individual to Group Privacy in Big Data Analytics” (2017) を発表している。

Brent Mittelstadt (2017) “From Individual to Group Privacy in Big Data Analytics”. *Philosophy and Technology*. 30, 475-494.

<https://link.springer.com/article/10.1007%2Fs13347-017-0253-7> (最終閲覧日 2020/6/29)



った問題ではない。しかしビッグデータの場合には、データ量が多いため喪失される規模も大きく、しかも喪失されたものを特定することも難しく固有の困難さがある。それゆえ、問題がより深刻となりうることを認識し、さらに考察される必要がある。

### ③信認関係(Fiduciary Relationship)

ビッグデータはデータ主体とデータ管理者の関係を変質させる可能性がある。例えば、医療においては、ビッグデータに信頼をおくことで、患者への理解や共感がおろそかになる可能性がある。こうした関係性の変質にも敏感である必要がある。

### ④学術的活動と商業活動

今後ビッグデータの問題を考える際には、それが学術的な活動なのか、商業的活動なのか、区別するべきである。というのも、両者は目的が異なりデータ主体にとっても重要な区別であるうえ、一般に商業利用は監視の体制が不十分であるという傾向があるからだ。

### ⑤知的財産の所有権

データ主体には、自分のデータによって生み出された利益に対する所有権も認められるべきであるという議論はさらに深められるべきであろう。人々から無報酬で収集したデータの分析により、相当な利益が生み出されることを考えると、オンラインでの購買やサービスの利用などで生み出された個人データに対し、すべての権利を放棄するような現在の所有権の枠組みは再考される必要がある。

### ⑥データアクセス権

ビッグデータでは、その理解のために、インフラ、スキル、知識が求められるが、これらをデータ主体が備えるのは現実的でない。意味のあるデータアクセス権を保障するためには、データ管理者にその実行責任を課すことになる。しかし、この要求についてはさらなる正当化が必要であるし、合理的なアクセス権についても、企業の商業的利益とバランスしながら議論する必要がある。また、個人情報保護法の改正を視野に入れた議論も今後求められるだろう。

## 結論

新たな技術の登場や、科学の進歩が起きたとき、それによって期待される利益が過度に強調される傾向にある。同様のことがビッグデータについても指摘できる。こうした傾向は、人々、マスメディア、そして研究者に対しても、問題のある仕方で影響を与える可能性がある。本稿では、ビッグデータの倫理的問題を取り上げてきた。これを通して、既存の楽観的態度を改める必要性を指摘することを試みた。とりわけ、データの取扱いに一層の慎重さが求められる生物医学の領域では尚更である。本稿が、ビッグデータ時代に限ることなく、生

物医学における倫理的な、そして責任ある発展に寄与することを望んでいる。

(荻野 琴)