AI ネットワーク社会推進会議

AI 経済検討会

第９回　議事概要

１．日時

令和２年１月 31 日（金）13：00～15：00

２．場所

三田共用会議所第４特別会議室

３．出席者

（１） 構成員

岩田座長、安宅構成員、石井構成員、喜連川構成員、久保田構成員、桑津構成員、実積構成員、杉山構成員、立本構成員、原田構成員、山本構成員

（２） 総務省

磯情報流通行政局地域通信振興課長、岡本国際戦略局多国間経済室長、井上情報通信政策研究所長、山田情報通信政策研究所調査研究部長、飯島情報通信政策研究所調査研究部主任研究官

（３） オブザーバー

内閣官房、消費者庁、個人情報保護委員会、経済産業省、理化学研究所、科学技術振興機構

４．議事概要

（１）事務局からの説明

事務局より、資料１及び資料２に基づき、「AI ネットワーク社会推進会議 ＡＩ経済検討会 運営方針（改定版）」及び「今後の検討の進め方（案）」について説明が行われた。

（２）中林氏からの説明

中林ヤマトホールディングス株式会社社長室デジタルイノベーション担当シニアマネージャーより、資料３に基づき説明が行われた。

（３）意見交換

【杉山構成員】

・　AIを使って、データトランスフォーメーションをすることで、収益拡大は企業としていろいろやっていると思う。アカデミアの立場からすると、サステイナビリティーへの取組に非常に興味がある。御社としての考えを伺いたい。

【中林氏】

・　一番わかりやすい例は、CO_2についての課題だ。EV化を進めるなどCO_2排出を抑制することを考えている。

【杉山構成員】

・　ドライバー不足という問題が大きく出ているかと思う。これに関して何か取組などしているか。

【中林氏】

・　非効率な部分がドライバー不足と見えているところもある。まずは今何が起こっているか可視化し、必要な人員配置をするという仮説を立て、進めている。少しずつ効果が出ている。必要な人員配置が分かれば、ドライバー不足ではない部分が分かる。足りない部分もあるので、そこは人ではなく違った配達方法があると考えている。その両面でドライバー不足は解決する方向で取り組んでいる。

【喜連川構成員】

・　過去にクロノゲートを紹介いただいた。当時、18億個のODデータを使ってこれからチャレンジする、データファーストだと言われた。そのときにやはりリーガルイシュー

やデータガバナンスという課題が出てくる。物流は、センシティビティーの高い情報だ。どの位の情報にすれば、サードパートシェアリングができるのか。その辺をどうされるのか。また、フィジカルインターネットの視点からみて、御社は現状、どのような感じなのか。

【中林氏】

- そういう戦略をつくって実行しろとアポイントされている。1つは、ようやく構造改革で事業戦略を発表できた。その中で、データファーストの戦略が組み込めたので、今、そこを着手している。
- 44年前のビジネスモデルのままの契約条項、契約書、宅急便の約款であるため、個人情報保護法を全く加味していない。そのため専門のリーガルチームを組成した。まずはデータファーストで蓄積できる状態をつくる。蓄積後に、それをどうサードパーティーと交換、場に出してミックスするのかのフェーズになる。まずは法的な問題をクリアしながら蓄積することに注力している。
- フィジカルリソースデータのサイバーへの落とし込み方については、アーキテクチャーのデザインを作っているところである。仮説を立てながら物理的なアーキテクチャー、地区センターとかベースと呼ばれている中継地をどこに置いたら一番効率がいいかというデザインになるが、まだデータがなくできないので、そのデータを蓄積し、ある程度蓄積したところで、センターとかベースのリアルな物理的な配置を決めながら、その上でトラックやドローンも視野に入れどう荷物を流せばよいかなど、配送リソースのコントロールにつなげていきたい。

【立本構成員】

- 従来の契約だと、個人情報を蓄積するのは難しく、そこを見直し、蓄積できるようにしたという話だが、ポリシーは具体的にどういうポリシーになるのか。つまり、企業の中には、個人情報を蓄積するのはリスクが高すぎて、現場は嫌だという企業もある。一方、そこをやっていかないと駄目ではないかという話もある。その辺の考えを聞きたい。

【中林氏】

・　契約内容を見直しながら、既にあるデータや取れそうなデータの洗い出しをしているところ。その中で個人情報については、運んでいるものも含めるとかなりセンシティブな内容になるので、取り得る情報を並べながら、どこまでをとりにいくかを検討する。プロコンはリーガルメンバーと検討し、幾つかのオプションをつくり、社長含めて経営会議に諮る方向で進めている。

・　プロコンは、個人情報を自社で抱えるリスクもあるが、我々の顧客に対してベネフィットを提供することで、顧客にカンファタブルな状態でデータを蓄積することもできる。インセンティブ設計もあわせて考えなければいけない。データを貯める、活用するとリピュテーションリスクも出てくる。これに対してはどうリピュテーションリスクをヘッジしながら良いサービスを提供していくか、さらにデータをどううまく取っていくかを、多様な軸で検討を始めている。

【安宅構成員】

・　資料３の28ページのベン図について。いわゆるCTOが一番下にあり、その上にCDOたちが３つある。要は、今まで外注してきたシステム構築をデータ・ドリブンな会社にするために内製化しようとしなければ回せないという話だ。これは非常に本質的で、経団連でAI-Ready化ガイドラインを議論したときも、99.9%の日本の企業はシステムインテグレーターにここを依存しているというのが主な議論だった。この内製化をどうやって加速化するかは、日本国の経済発展ために非常に重要な鍵になる話だ。

・　この人づくりが並大抵ではない。Slerから人を派遣してもらえば済む問題ではない。この上のビッグデータ、そしてAI的に実装できる人はそもそもいない可能性が高い。これをどうやって作っていくかというのは国家的課題として例示されたと認識した。

【中林氏】

・　補足すると、今のチームはデジタルの組織、300人の組織だ。CDO、CTOと分けずに、自分がその役割だと思っている。CTDO的な、テクノロジーもデータも含めて自分がコントロールする。そのことで４つの枠が自分の下でコントロールできる。組織もそういうデザインをしている。アジャイル開発チーム及びデータサイエンスチーム、あとはビジネスとのコミュニケーションをする企画チームを作って、そこに人を集めなが

らコラボレーションさせていくように進めている。

【安宅構成員】

・　だとすると４つに組織を分けなくても良いかもしれないが、ヤフーというデータが多い組織にいると、システムエンジニアリングの組織だけで1,500人から2,000人おり、さらにデータエンジニアリングとデータサイエンスを担う組織で1,000人いる。それぞれ独立な巨大組織であるので、割らざるを得ないときは必ず来る。また、メガトラフィックをさばくのは並大抵のエンジニアリング技術ではないので、これもどこかで割らざるを得なくなる。

【原田構成員】

・　大学に求める人材はどういうところにあるのかを、説明された観点から教えていただきたい。

【中林氏】

・　１つは、ある程度数学的な素養を持ってデータが使える人間が必要だ。これは、データサイエンス学部等で取り組まれているので、一定程度育ってくる。一番足りないのは、クラウドなどのインフラマネジメントから上のシステムエンジニアリング、又はアプリケーションをアジャイルに開発するスキルセットで、今の大学教育には全くない。そのあたりの人材が圧倒的に足りない。データサイエンスは育てられるという感触は得ているが、エンジニアはまだこれからのチャレンジだ。ハイトラフィックだとかなりのスキルセットを要求され、しかも可用性も求められる。そういったデザイン又は実装ができる人材が事業者側から見て足りない。

【山本構成員】

・　事業会社の中で人材を再教育するところで、ビジネスとデータサイエンスの両方を理解する人材を作ることが大事だというメッセージだった。マネージャークラスで再教育がどれぐらい可能なのか、あるいは年齢層、バックグラウンドなどでどうか。またマーケティングや事業戦略の担当の場合、ある程度数字を扱っているので入りやすい

が、例えばＨＲや人事だと、データサイエンスを理解しにくい。その場合、再教育は事業会社の中で可能なのか。

【中林氏】

・　これからのチャレンジである。前職では、アクチュアリーという狭い領域で数理統計に特化したメンバーがいたので、SQL、Pythonを使った機械学習を覚えさせると、現業が分かっているので、おもしろいアイデアとか実装の事業計画、企画を作ってくれるところまでの効果は見えている。現職では、理系の大学院卒で現場で働いているメンバーがいるので、まずはそこから再教育をし、組み込めるかどうかというチャレンジをしている。

【石井構成員】

・　１点目の質問として、個人情報を蓄積していくに当たって、どこからどこに何を運んでいるかが非常に機微性のある情報ではないかという話があったが、配送に関しては委託元からの契約上の縛りなどがあるのではないかと思ったが、これまでとこれからどうしていきたいのかについて伺いたい。
・　２点目の質問として、例えば、情報を配送リソースの効率化に使うなどの目的で使う場合は、個人と紐づくような形で蓄積する必要があるのか。管理していく情報の項目として、個人に紐づく可能性をどのように考えるかという点について伺いたい。

【中林氏】

・　１点目は、44年前のモデルを使い続けているので、契約書とか配送行為のみに使っており、しかも終わったら捨てるという形になっているのが現状。契約書の中身を見直す過程で、委託をされる側とする側がwin－winのビジネスモデルを考えている。
・　２点目は、とり得る情報を全部並べて、どこまでを取っていけばいいか３パターンから５パターンぐらい用意している。それぞれリスクやメリットもあるので考えられるところを並べて経営に諮っていこうと考えている。
・　現在はそもそも何を運んでいるか分かっていない。

【岩田座長】

・　例えば、再配達をミニマムにしようと思うと、個人情報と結び付けて今在宅か否かがリアルタイムで分かれば良い。在宅しているのをどう判断するかというと、電気の使用の有無だ。電力会社とデータを共有できれば、再配達のコストが下がる。一方、個人が在宅か否かが分かってしまうということは、個人情報を使わざるを得ない。消費者にはプライバシーと便利さのトレードオフがあり、それは、消費者自身が選ぶということにならざるを得ない。

・　１点目の質問として、長期的にはプラットフォーム化したい、物流のプラットフォームを最終的には作りたいとのことだが、そのプラットフォームはどういうものなのか。例えば、アマゾンは巨大なプラットフォーム企業だが、アマゾンと比べてヤマトのプラットフォームはどこが違い、どこに優位性があるのか。また、アマゾンと共存ができるのか、共存はできないで食うか食われるかになるという問題なのか。

・　２点目の質問として、最終的には価値を高めないといけないが、プラットフォームを提供するときに、収集したデータでどうやってマネタイズするかというのが次のビッグクエスチョンになるが、ヤマトの場合はどこが価値創造の目玉と考えているのか。アマゾンは本屋というｅコマースから始まり、エコシステムを拡大してあのように成長した。

・　３点目の質問として、サステイナビリティとこの事業をどう結び付けるのか。自分はエネルギーを一番使っているドライバー、車ではないかと考えている。ここをどのぐらいクリーンにするか。例えば、ドイツ・ポストは電動車を自分でつくり、全部１００％電動の車を使ってやっている。そのときに、使う電気がクリーンでないと具合が悪い。全部クリーンの電力を使えれば、配達に伴うエミッションはゼロになるが、どの辺までを射程に置いているのか。

・　４点目の質問として、必修科目で、前提条件の中にスクリプト系の言語で、Pythonが望ましいというが、Pythonのレベルまで、社員のどのぐらいの人たちが使えるのが望ましいか。

【中林氏】

・　１・２点目は関連する。プラットフォームとデータから価値を出すことはやり過ぎだ

と思っている。アマゾンもグーグルも、自分の現業を回していたらデータがたまったものと考えている。同じように我々も、年間18億個運んでいる荷物のデータを徹底的にデジタル化し切って、それを蓄積することにまずはフォーカスしている。利用者の個人情報を取るかどうかは別にして、個人と法人の利用者情報と荷物又は提供サービスの情報やそれを動かしているリソース（５万人のセールスドライバー、５万台の車、7,000の営業拠点、22万店の取扱店（クリーニング店、たばこ屋を含む））といったフィジカルなデータをまずはどう集めることでかなり事業のコスト効率が良くなるのは見えている。そこをやり切った後に、３つのフェーズと言ったが、まずはデータファースト、トランスフォーメーションのフェーズでデータを取り切るというところが第一義だ。その後、溜まってきたデータを見ながら、そこから何が生まれてくるかがイノベーションのフェーズになる。

・　３点目については、自社で５万台の車を持っているほか、そのメンテナンスのために、ヤマトオートワークスというメンテナンス会社を１つ持っている。ヤマトオートワークスは自社の５万台とアウトソースも受けている外の車５万台で10万台の車をメンテナンスできるような機能を持っており、日本のシェア10%を持っている。そこで、ヤマトオートワークスを使って車をアセンブルできるのではないかという仮説を作っている。量産までするかどうかは別にしても、10%のシェアを取るための車のデザイン等、試作品までは作って走らせられるという仮説を持って、内製化もオプションの１つに考えている。今、日本のシェア10%の車を全てＥＶ化するというところも目指している。

・　４点目については、組織の再編を行っているところ。どのレイヤまで、何を、どう教育するかという設計をやっている。資料３の14ページで示したCenter of Excellenceの中にPythonのコーディングができる人は集約してしまい、一方、現場は機械学習で予測された荷物の量をもとにリソースコントロールするというオペレーションを落とすので、Pythonは必要ないと思っている。ただ、その機械学習が何であるか、それがどうやってできているかという理解はさせようと思っている。Pythonを把握しているCenter of Excellence の人材は数十人規模で抱えようと考えている。一方、最近、オートML、データロボットなどで自動化されつつあるので、そこはある程度、機械化できると見ている。

（５）田丸氏からの説明

田丸日本マイクロソフト株式会社業務執行役員ナショナルテクノロジーオフィサーより、資料４に基づき説明が行われた。

（６）意見交換

【実積構成員】

・　データの取扱に関して、企業が追求する価値というものと、社会の価値というものの整合性というのはどういうふうに担保していけばいいとお考えなのか。特に公平性という問題だと、企業が考える公平性と、それぞれの国における公平性の概念とかが少し違うところがあるのではないかという懸念をしている。AI倫理審査ボードで公平性の基準をお持ちだと思うが、それが企業がビジネスをしている市場における公平性と一致していることを担保する仕組みというか、少なくとも社会のほうから見て、マイクロソフトが適用している公平性の基準というものが、受け入れる枠内にいるということをどのようにして担保するのが望ましいとお考えなのか。

・　「AIデータ活用コンソーシアム」に関与している企業というのはそれなりの責任感を持っていると思うが、今後、新しい企業が入ってきた場合に、その企業が提供しているデータの利活用の方針というものが正しいものなのかどうか、適切なのかどうかというものを確認する仕組みについて、何かアイデアをお持ちであればお伺いしたい。製品、サービスに対して、社会の規範からずれているかもしれないというのを見つけるシステムというのが今後必要になってくると思っている。

【田丸氏】

・　データの価値は、それを必要とする側によって大きく異なってくる。一方で、企業が持っているデータをそもそもオープンにするべきデータなのかどうなのかという議論もある。データの価値についてはマーケットが決めていくものであると考えている。ただし、データのドメインなど様々な要素がある中で、流通のための基盤が現状存在をしていないというのが我々の認識である。「AIデータ活用コンソーシアム」ではデータ流通の基盤を作ろうというのが１つのゴールになっている。流通であれば、ユーザーは、

企業の評価や価格など様々な属性情報から、どこから買う、どれを買うのかというのを決めていると思うが、最終的にはデータもそうあるべきと考えている。個人情報など機微なデータをどう考えるのかということはあるが、一般的に必要、価値あるデータというのは、必ずしも個人情報に関係するような機微なデータとは限らず、それ以外の情報は多くあるので、そういったものを流通できる場を設けるということが重要である。

・　「AIデータ活用コンソーシアム」の会員というのは、データ基盤の構築や契約のワーキンググループも含めて、一緒に作業していただくというところでの会員、研究者の先生方であり、実際にデータを流通させるために、会員である必要はない。無償のデータも有償のデータも、提供条件も含めて、データホルダーが決める仕組みになっている。

・　公平性をどう考えるのかについては、マイクロソフトの中では、ご指摘のとおり、立場、シーン、さまざまな要素によって公平性のポジションが異なる。マイクロソフトの中の人間だけでは、全ての考えるべきケースを洗い出し、また、適正に評価できるとは考えておらず、適切な有識者に依頼をして、検討委員会のような場で、第三者の意見も反映させつつ検討することを行っている。それによって、社内の目だけではなくて、多分に社外からの評価を反映させることによって、可能な限り公平、公正な判断ができるような努力をしている。

・　「AIデータ活用コンソーシアム」の会員になることについては、特に制限を設けておらず、データ流通というところにおいては、別に会員である必要性は全くなく、特に制約を設けている取組ではない。

【実積構成員】

・　最後の利用者から見て、データの取り扱いがおかしかったり、AIの学習が公平でなかったりすることを見つける仕組みというのは可能か。車だと事故が起きるとまずいというのは分かるが、AIの場合は外に結果がないと分からない。今までの新聞に出てくるものも、特定の専門家、あるいは学術研究者がデータを壊してみてどうかを判断する経過になっている。一般の利用者がおかしさに気づくことは、この分野では可能なのか。それとも、そこは特定の専門家集団が活躍する場になるのか。

【田丸氏】

・　非常に難しい領域だ。AIに限らず、ルール、ロジックで実装されているシステムについても同様だ。その中身の理屈がAIの場合には説明可能性が低いだけのことだ。その結果について、リーズナブルなのかそうでないのかは、特にこのAIに限ったことではないのではないか。それを適切に評価できるのかは、従来のシステムと議論としては大きく変わらない。一方、そのデータをどう考えるのかは難しい。難しいが故にいろいろな研究者がxAIということで取り組まれている状況だ。ちなみにコンソーシアム自体は、主に知財・契約の領域、あとはデータ基盤の構築で、その結果のインファレンスのエンジンの品質がどうなのかというところまでは、取組のスコープに入れていない。

【岩田座長】

・　１点目の質問として、この「AIデータ活用コンソーシアム」と「データ・トランスファー・プロジェクト」との関係はどうなっているのか。また、データの共有化と言うとき、オープンAPIというのが特に金融サービスでは大きくなっている。「AIデータ活用コンソーシアム」でデータを共有するとき、APIのところでオープンにする方式を考えているのか。また、技術としてブロックチェーンは使われるのか。
・　２点目の質問として、プライバシーと関係するが、資料４の23ページでデータドメイン、知財、商流、来歴、品質、契約となっていて、来歴のところで「説明可能性とトレーサビリティー」というのがあり、トレーサビリティーがプライバシーとぶつかることが起こり得ると考えている。プライバシーの保護とトレーサビリティーの関係をどのように整理されているか。
・　３点目の質問として、外部データ活用の統制がより重要になるというのはそのとおりだが、統制は誰がやるのか。「AIデータ活用コンソーシアム」がやるのか、それとも政府部門がやるのか。

【田丸氏】

・　１点目について、「データ・トランスファー・プロジェクト」と、「AIデータ活用コンソーシアム」の取組は一切関係ない。「データ・トランスファー・プロジェクト」は米国で起きている活動であり、基本的にはデータポータビリティーの視点で始まった活動である。一方、「AIデータ活用コンソーシアム」の背景については、特に日本にお

いて、データの先制取得者と最終的な利用者が同一のケースばかりではないことにある。有益なデータを持っている企業のデータをいかに流通させ、最終的に国内の研究活動、経済に資することができるかという視点で活動している。

・APIについては、APIで考えるのか、データフォーマットで考えるのかということになる。「AIデータ活用コンソーシアム」のデータ基盤は、特に取り扱うデータの種類については制限を設けていない。これは、産業によっては、特定のデータについてフォーマットが標準化されているケースもあるし、APIでのデータのやりとりが標準化されているような領域もある。標準化されているものは、APIの隙間というか、使用を含めて登録ができる。データフォーマットの情報を含めて、データカタログに登録をすることで、利用者がそれを見て購入、利用するような形での実装を進めている。

・ブロックチェーンについては、契約の手続と、ラベルデータ、アノテーションで活用を検討している。例えば、画像のデータを売るベンダーや個人、実際にアノテーション、ラベルデータを作り販売するベンダーについて、そのエンティティーが必ずしも同じではなく、別々のケースが増えてくるだろうと考えている。画像データを買う側からすると、自分が欲しいラベルデータ、仕様に基づくラベルデータが既にマーケットにあれば、そのラベルデータと、その元画像の両者を買いたいと思う。一方、画像のデータオーナーからすると、自分のラベルデータを色々な業者が売ってくれれば、自分の画像データが売れる可能性が上がる。また、契約手続の中でも、契約の親子関係が出てくるので、その親子関係をどうクラウド上のサービスとして扱うのかという点では、仕組みとしてそこにブロックチェーンを使うことを検討している。また、アノテーションのラベルデータについても、説明責任を果たすといったところでデータを署名する技術の選択肢としてブロックチェーンもあるが、現時点では明確ではない。

・　２点目について、トレーサビリティーは、説明責任を果たせるかどうかがポイントであり、製造物責任を含めて説明しなくてはいけない際に、証明できるかという文脈でのトレーサビリティーであるため、必ずしも流通の段階において、データ開示されるべきものではないと考えている。その個人情報の扱い等は、実際に事故が起きたときに、それをどう説明するのかということになるので、違うスコープの話になる。

・　３点目について、企業にとっては出所不明な外部データを使うこと自体、リスクが高まってくるので、データ活用統制の説明の中で、悪意を持ったラベルデータが混入した

データを知らずに使ったり、データの偏りがどうなっているかも分からないデータを使ったりした結果、でき上がった製品、品質に非常に問題が生じるといったことに陥らないために、外部データを使うときの活用基準をしっかり企業として考えていかなくてはいけないという意味で、企業側の外部データ活用の統制という説明をしている。

【杉山構成員】

・　最近、大手の国際的なIT系企業は、個人データを集めませんということを売りにしているようなケースが増えてきている。個人的な感覚としては不公平を生むのではないかと感じている。それは新規参入する会社はまだデータを持っていないのに、これからはデータを取ってはいけない時代ということになると、例えば、核実験を十分やった国がデータをたっぷりためておいて、これからは核実験をしてはいけないと言っているのと、雰囲気が似ている。マイクロソフトがどちら側かというのはまたちょっと別の話かもしれない。データを、個人情報を取ることがあまり良くないという雰囲気になってきているのは事実かと思う。それが行き過ぎると、別の意味での格差、参入障壁になる。何かそういうことに関してご示唆することがあるか。

【田丸氏】

・　マイクロソフトは、ホットメールを始めたころから、基本的に顧客データは研究開発に使わないと明言している。米国だと、ホットメールの時代は、特にメディアは、ホットメールしか使わなかった。また、他社から来たリサーチだと、マイクロソフトリサーチのユーザーデータには全くアクセスができず、研究者にとっては全然データがないとよく文句言われる。
・　一方、マイクロソフトは、これまでもビジネスを通してデータを集めるよりも、例えば機械翻訳の対訳コーパスでもそうだが、いろんな対訳コーパスを持ち寄ってそれを相互に活用する取組がある。そういった翻訳ベンダーや、色々な企業のコミュニティーの中でデータを集める、共有する、あとは実際にコストをかけてデータを作るということを行ってきている。そういった意味では、マイクロソフトはビジネス活動を通してデータを集めるよりも、自社で作るということをしている企業だ。
・　データを既に持っているところが有利なのではないかという指摘は、個人的にはそ

のとおりだと思う。EUだとGDPR等あるが、対象とするデータがなくても周辺データで、その中核データを浮き上がらせる、ある程度推定することができるという考え方が高まっている。逆を言えば、個人識別可能なデータを削除したとしても、周辺データを十分に持っていれば、無いデータを復活させるということが技術的にも可能になってきている。それを考えると、既にデータを持っている側、持っていない側というギャップはより広がる傾向にある。そういう研究を深め、ファクトを見ると、そういう傾向が明らかにあるのではないか。

【原田構成員】

- データの流通に関して、例えば、コンピュータービジョンとか、ビジョンリコミッションという画像認識がある。成功したのがちょうど2012年度で、ディープニューラルネットワークがすごいという話になった。ディープニューラルネットも確かにすごいが、その裏にImageNetというデータセットがある。成功の半分はそのデータセットであろうということだ。実は画像認識の歴史は50年ぐらいあって、どういうアルゴリズムとか理論を作っていけばいいかだけではなくて、その裏にはどういうデータセットを作っていけばいいのかという歴史もある。データセットをずっと作っていく、データセットバイアスをどうやってなくしていくかという歴史で、ようやく完成したのがImageNetというデータセットだ。それを使うことによって、今のディープニューラルネットワークの大成功がある。
- 画像認識をやるとき、データセットを作ってアルゴリズムに読み込ませ、それでアルゴリズムの成果を見る。そのときデータセットバイアスの出方を見て、もうちょっと違うデータセットでバイアスをなくしたデータセットを作ることで、いいデータができてきた。
- データ流通で、データセットバイアスがあるからこのデータを何とか使えるようにしてくれとか、そういうフィードバック、循環で回っている。バイアスが少なくなったものを育てていくような枠組みになっているとすばらしいと思っている。その点について意見を聞きたい。

【田丸氏】

・　「AIデータ活用コンソーシアム」が直接的にバイアスをなくしていくことに、直接に寄与できるとは考えていない。一方、コンソーシアムはデータカタログという言い方をしているが、カタログの中で、データの偏りも含めてメタ情報として可能な限りデータを持っているようにしている。もう一つは、何をもってバイアスがあると考えるのかという評価手法についても今後検討する。それに基づいた情報をカタログデータとして付加して、購入する側は、それを知った上で買うことができるようにすることは議論している。

【原田構成員】

・　ベンチマークテストみたいのが10種類、100種類あって、それに入れたときに、このベンチマークの正確性がどれくらいかがざっと出てくるような形か。

【田丸氏】

・　その通り。当然全てのデータについてそれをやるということではない。当然そういう情報があるデータもあればないデータもある。来歴情報がしっかりしているデータもあればないデータもある。基盤としては、そういう情報があるデータやないデータは、当然マーケットでは値段が自然に変わっていく。そういう情報を含め、しっかりしているデータはそれなりに高い値段、そういう情報でないと買えないユーザーからすれば、値段が多少高くても買うだろう。単純に、教育に少し使うだけということであればそこまで求めないだろう。そういった意味で、ニーズと実際のデータの有り様というところで価格は変わっていく、決まっていくものではないか。
・　アノテーションスペックも目的によって異なってくる。私自身がやっていたとき、画像データセットのときにも、マイクロソフトの中では、どういう基準でという議論はあった。ラベリングについても、用途によってアノテーションスペックが異なるので、例えば防犯目的に何かリコグニッションをということであれば、体のパーツの一部だけ映っていても全部ラベリングするだろう。自動ドアで自動的に人が全身映ったときだけディテクションしたいといった場合には、全身映っているところだけ動く設定するかもしれない。「AIデータ活用コンソーシアム」の基盤としては、ラベルデータについても、どういうラベリングスペックの下にこのラベリングをしているのか、セットで情

報をカタログ上で参照できるようにすることによって、買う側はそれを理解した上で買えるようにしようと考えている。

以上