1E1-OS-24a-3

人工知能の製造物責任とリスクに関する試論 トロッコ問題を例として

Discussion about Product Liablity and Risk Management of Products with Artificial Intelligence Using Trolley Problem as an Example

吉岡 真治 *1
Masaharu YOSHIOKA

*1北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In this paper, we discuss issues related to product liability and risk management of products with artificial intelligence(AI) in the open world. Since products with AI behave autonomously, there is a discussion about liability of such product based on their decision. However, it makes difficult for the manufacturing companies to sell such products. In this paper, we introduce the framework for sharing the responsibility of the decision with manufacturing companies and the users based on how they recognize risk of using the products. This framework is investigated by using trolley problem as an example.

1. はじめに

現在、自動運転に関する技術開発に代表されるように、人工知能を搭載して、自律的に行動する製品の開発が行われている。本試論では、このような自律的な製品がなにか問題を起こしたときの、その責任の所在についての議論を行う。

ここでは、その議論の出発点として、社会的にも受容されている運転手の補助装置としての自動運転技術に関する製造物責任の議論、飛行機などの特殊用途での自動運転技術に関する現状を考察すると共に、車の自動運転技術を社会的に受容させる際に起こる問題点について議論する。

特に、自律的な製品が行う判断に大きな影響を与えると考えられる製品の製造者、利用者、その製品が利用される環境に存在する他者という異なるステークホルダーを想定し、各々の責任という観点から、この問題点について議論する。特に、各々のステークホルダーの判断について、判断に伴い許容したリスクという観点から各々の責任を議論する枠組を示す。

この枠組の具体的な適用事例として、近年の自律的な人工知能に関する倫理的な側面に関する議論で良く用いられる倫理学における思考実験であったトロッコ問題 [Foot 67] を例にとり、具体的な議論を行う。

2. 人工物と製造物責任

製造物責任法とは、消費者保護の観点から、利用者の不注意ではなく通常の利用を行った際に起こる問題を、製造者の責任として問うための法律である。しかし、通常の利用という定義が明確でないため、説明書に記載していない禁止事項以外の項目の利用により生じた問題が、例え、一般的には通常の利用と考えられなくても、製造者に責任を問うという法律的な判断がなされることになった。そのため、製造者は、通常の利用に関する禁止事項を幅広く設定することで、訴訟のリスクを回避するようになってきている。このことは、製造者の想定外の状況で利用した場合の責任は、基本的に利用者に存在することを意味している。

このような製造物責任の考え方をベースに自動運転について考えてみる。上記のような制約条件を与えることができる自動運転システムの応用分野としては、工場のようなクローズな環境下における自動運搬装置の自動運転システムが考えられる。このような応用分野では、工場における環境を整備することや、人間とのかかわり方を明確にすることにより、製造者と利用者の間での責任分担を明確にした形での人工物の利用が可能となる。

しかし、一般の道路環境の様なオープンな環境では、この様な制約を事前に満すように環境を整備することは困難である。そのため、現時点の自動運転に関わる技術は、通常の道路環境という限定的な状況でのみ運転に関わる操作を補助し、最終的な操作は人間が対応するという形で提供されることになっている。近年の自動運転技術の進展を踏まえ、米国運輸省は、SAE J3016 において、車の制御、対象物の認知、例外的な事象への対応、全体のプランニングについての運転者とシステムの分担の仕方に応じた自動運転技術のレベル分けを行っている。この分類に従うと、現時点で提供されている技術は、自動運転ではなく、運転者の補助装置であり、最終的な判断は、運転者が行うことを前提とした技術となっている。この枠組では、事故などが起こった場合には、基本的には、その責任は、運転者に存在することになっており、この様な装備を備えたテスラの車が起こした 2016 年の死亡事故に関する報告 *1 においても、車の認識能力や操作の是非ではなく、運転者にどの様に注意喚起を行っていたかという点が議論され、現時点で車に問題はなかったという結論となっている。

同様の機能と責任の分担は、飛行機の自動運転についても行われているが、飛行機の場合は、運転者のほとんどが、職業パイロットであり、例外的な条件の発生が、あらかじめある程度予想できるような状況であるため、一定の緊張感のもとに実現可能となっていると考えている。これに対し、テスラの車の事故に代表されるように、車の運転のように、利用者の多くが一般の利用者にとって、多少の注意喚起があったとしても、例外的な状況がいつ起きるかわからない環境において、この様な緊張感を維持し続けることは困難だと考えられる。また、運転補助技術が高度になればなるほど、運転者が車を直接操作をする機会は減少し、さらに注意を持続し続けることが困難となってくることが予想される。

連絡先: 吉岡 真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7107, yoshioka@ist.hokudai.ac.jp

*1 https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.pdf

一方で、例外的な事象についてもシステムが対応するという状況になると、製造者に関する責任が生じてくることが想定される。このレベルになると、倫理的な判断に関する思考実験として用いられているトロッコ問題 [Foot 67] を題材として、人工知能の判断の倫理的な側面などの議論が必要となる。トロッコ問題とは、制御がきかなくなったトロッコが線路の切り替えポイントにさしかかろうとしている時に、どちらのポイントに切り替えても何らかの被害が出るという状況下での判断を問う問題である。人数の違い、年齢の違い（若者、老人）、社会的な立場の違い（大統領、一般人）などの様々な条件を設定して、判断をさせることにより、その判断をする際の倫理的な判断基準を議論することが可能となる。また、インターネット上で、トロッコ問題を自動運転の問題に置き換えた際にその判断の妥当性を問うためのサイトが公開されると共に、どの様な自動運転システムを作るべきかが議論されている [Bonnefon 16]。しかし、この中では、ユーザーがどの様な判断をするタイプの製品を選びたいと思うのかといった議論に留まっている。

3. 複数のステークホルダーを想定した責任分担の枠組

3.1 トロッコ問題を例とした責任とリスクに関する議論

前節で議論したように、現状の運転者の補助装置や自動運転システムでは、不測の事態への最終的な判断を運転者に委ねることで、基本的な責任を運転者に委ねる一方で、不測の事態への対応を車の側が引き受けた途端に、その倫理的な側面を含む形で責任を引き受けるという極端な議論になってしまっている。

この様な極端な状況を解消するためには、製造者と利用者の間での適切な責任の分担を行う必要がある。また、トロッコ問題のように、製品が利用される環境にいる他の人がその判断に影響を与える可能性が存在する。これらの状況を考慮して、本稿では、このような議論の問題点について、各々のステークホルダー (利用者、製造者、環境に存在する人 (例えば、トロッコ問題で事故にあう可能性のある人など)) が負うべき責任について議論する。

まず、一般的な議論を始める前に、トロッコの製造者、トロッコの運用責任者、事故にあう可能性のある人（被害者）という 3 タイプのステークホルダーのみが存在するという最も簡単な条件におけるトロッコ問題を題材として、リスクと責任に関する議論を行う。

トロッコ問題の議論は、問題を単純化して、ポイントの切り替えを行うトロッコの運用責任者の判断にのみ注目した議論となっている。しかし、この問題を実問題として捉える場合には、トロッコが制御不能になったということで、トロッコの製造者に関する責任も存在する。また、被害者にも、そのような場所にいた責任があるとも考えられる。このような責任について議論するためには、インターネット上での調査 [Bonnefon 16] でも行っているような単純な問いかけでは不十分である。例えば、被害者の責任について考えるためには、その場所の危険性をどのように理解していたかが問題になる。例えば、トロッコの運用責任者がその場所は、トロッコに問題があった時に、意図的に退避させる場所としてアナウンスされている場所なのであれば、被害者はそのリスクを理解していたことになり、一定の責任が生じると考えられる。トロッコの製造者にしても、制御不能になったのが、突然の事態なのか、あらかじめ異常を知らせていたのかによっても、責任の重さが変わる。トロッコの運用責任者については、トロッコの問題をどの程度、事前に把握していたのか、退避線に入っても止まれないようなスピードで運用したことに関するリスクなどをどのように理解していたかが問題になる。

このように、単純なトロッコ問題であっても、各々のステークホルダーの問題の認識によって、その責任の割合は大きく異なり、単純にトロッコのポイントの切り替えの局面の判断のみに注目することは、議論としてあまり適切ではないと考える。実際に、同じような交通事故が起きた場合でも、加害者や被害者の事故に至った経緯などを考慮して、各々の過失割合が決定されていることもこの議論の社会的な側面からの妥当性を支持していると考えている。

では、ここで、運用責任者の立場に立って、この事故に関する責任を取るリスクを回避する方法を考えてみる。一番簡単な方法は、退避線の長さや、非常停止装置の性能を向上させるとともに、トロッコがその退避線で十分に止まることができる速度に最高速度を設定することにより、トロッコが、線路の外に飛び出さない環境を作るという方法である。これは、リスクを可能な限り回避するという観点から考えるとよい方法であるが、経済合理性の観点からは設備投資が過大となるとともに、輸送効率が下がり問題となる。よって、運用責任者は一定程度の事故のリスクを許容しながらの運用を行うこととなる。これは、他のステークホルダーにも同様の議論を行うことができ、事故が起こるということは、何らかの意味で、各々のステークホルダーがリスクを許容した結果として起きた結果であると考えられる。

このようなリスクは、複数のステークホルダーが連携することにより、個々のステークホルダーが行うリスク回避よりも効率的に回避できる可能性がある。例えば、トロッコが故障する可能性が高くなっていることをあらかじめ通知できたならば、トロッコの運行速度を減速するという運用や、退避線を使う可能性が高くなってきたときに、退避線の先にいる人々に、場所を移動するように促すといったシステムを構築することで、経済合理性を維持しながらリスクを低減することが可能となる。

3.2 リスク許容レベルを考慮した責任の分担

3.1 節で述べたように、オープン環境における不測の事態への対応を考える際に、単純なトロッコ問題の問いかけのように、他のステークホルダーの事情を考えずに議論することは、倫理観にのみ依拠することになり、人によって基準が違う [Bonnefon 16] という段階から抜け出すことが困難となる。

そこで、各々のステークホルダーの状況を考慮した中でのトロッコ問題への対応について考える。この時、人工知能は、可能な限り被害を食い止めるように努力するという前提をおくと、このトロッコ問題での選択は、各ステークホルダー間の被害に関する複数のパレート最適の解 *2 から、一定の基準で解を選ぶというタスクに相当する。

この様なパレート最適の解の中から判断を下すというタスクは、複数の異なる評価基準 (各々のステークホルダーの被害の最小化) を統合して、何らかの解を選ぶ基準が必要となる。本試論では、この基準として倫理観といった曖昧な概念ではなく、ステークホルダーがどの様なリスクの存在を許容していたのかといった情報を与えることで、同様の事故の場合の過失割合などに関するこれまでの判断といった情報から、責任分担の割合を考える方法を提案する。このような責任割合に関する情

*2 特定のステークホルダーの被害が他のステークホルダーの被害を増やさずに軽減できない様な解。具体的には、一方のステークホルダーの被害を減少させるためには、ポイントを切り替えるしか手段がなく、結果として、他のステークホルダーの被害を増加させるといった状況は、パレート最適の 2 つの解が存在すると考える。

報を用いると、ステークホルダーの責任割合と被害を対応させるといった基準を導入することや、他のステークホルダーへの弁済を最小化するような基準など、単純な倫理観ではない判断基準を導入することができ、人工知能は、その範囲の中で最善を尽くす (パレート最適な解を見つける) という対応が実現できる。

この様な枠組を設定した場合の製造者と利用者の責任分担について考える。前提として、製造者は利用者に対して、許容するリスクレベルに関する情報と、パレート最適の解の中から一つの解を選ぶための基準 (例えば、上記のような経済合理性を考えた基準だけでなく、利用者に深刻な被害が及ばない限り、他人に危害を与えないといった基準も選択可能とする。) を与えることとする。この場合に製造者は、人工知能が下す判断にたいして、与えられた基準に応じて最善を尽くしたかどうか (下した決断により得られた結果がパレート最適な解の一つであり、かつ、利用者の設定した基準にそった結果になっているかどうか) について責任を持つということになる。また、人工知能を含む製造物全体としては、その不測の事態に至った原因が、製造物の機能的な問題 (センサーがうまく働かなかった) にない場合には、責任は利用者にあると考える。結果として、製造者は、その判断の倫理的な問題に関する議論を避けることが出来るだけでなく、利用者の通常とは異なる利用 (社会通念上、想定してもおかしくないリスク許容レベル以上の利用:例えば、約束に遅れそうなので、スピードを通常よりあげて運転をする) をした場合に起きた不測の事態への責任について、その責任の多くを免れることとなる。結果として、現状の制約条件を与えた上での利用を想定した製造物に対する責任と同様の責任のみを引き受ける形で、オープンな環境で利用可能な製造物を社会に提供することができるようになると考えている。

3.3 社会通念上、存在を許容するリスク

ここで、許容すべきリスクに関する判断の全てを利用者に任せるという状況を考える。この場合に、リスクというものは回避すべきものという対応をする利用者を想定すると、ほとんどリスクを許容しない (例えば、近距離のセンサーで、人や他の車を関知した際には、その人や車が、突然、こちらに向かってきたり、車が急ブレーキをかけたとしても、余裕を持って停止できるスピード以上のスピードを出さない) といった設定も可能となる。しかし、この様な設定を行った場合には、制限速度とは関係なくゆっくりしか走ることが出来ない場合が多くなるため、限られた道路という資源の有効活用という観点からは問題となる。

この様な問題を避けるために、社会通念上、存在を許容するリスクという考え方を導入し、そのリスクについては、可能な限り全てのステークホルダーがその存在を理解し、そのリスクが顕在化しないように努力することを義務づける必要がある。具体的には、例えば、車の往来をある程度優先するような道 (国道などの主要道) では、ガードレールなどで車と歩行者の間を仕切ると共に、歩行者は、むやみに道を横切らない、といった現実の世界の交通ルールの遵守を徹底させることを想定する。この時、限りなくリスクを回避したいと考える利用者にとっても、歩道の歩行者の横を通り抜ける場合には、少なくとも接触事故という観点からは、歩行者が道に飛び出てくる可能性を 0(0 にはならないが、このリスクについては、社会通念上、歩行者の側がその責任を引き受ける) と考えて、リスクを回避したまま、通常通りのスピードでの運行が可能となる。逆に、この様な状況の道において、道を横切ろうとした歩行者が接触事故を起こした場合には、その歩行者がとるべき責任は、他の道を横切ろうとした場合に比較して、社会通念上、より重い責任が問われるということになる。

3.4 不測の事態の想定に関する分析

これまでの議論において、関係するステークホルダーが明確になった場合には、ステークホルダーが協力することにより、不測の事態を回避したり、その被害を低減できる可能性があることについて議論を行った。しかし、不測の事態は、そもそも関係するステークホルダーがあらかじめ確定しているような状況で起きるのではなく、その製品が利用される環境を分析することで、起こりうる不測の事態を想定し、関連するステークホルダーを確定させるという操作が必要となってくる。

一方で、起こりうる全ての不測の事態を網羅するということは、人工知能におけるフレーム問題と同様にほぼ不可能な問題である。また、ほとんど起こる可能性のない不測の事態についてもその可能性を考慮し続けることは、不要な計算を行い続けることになり、経済合理性の観点から問題となる。こちらについても、社会通念上、チェックを怠ることが問題だと考えられる事態 (起こる可能性や、その影響を考慮して作成) を共有することにより、製造者が最低限果たすべき責任を明確にすることが可能となると考えている。

付加的な機能としては、リスクの許容レベルについての情報 (特に社会通念上、存在が許容されるリスクレベルよりも高いレベルのリスクを許容しているステークホルダーが、影響を受ける範囲に存在するかどうか) が共有された場合には、単純なセンサー情報を用いるのとは別の形で、不測の事態の起こる可能性を計算するきっかけとする方法なども考えられる。

この様な不測の事態の予想とステークホルダーの発見に関する機能については、ある程度シミュレーションにより評価可能であることを考慮すると、例えば、事前に想定されるような様々な問題をどの時点で予想でき、それに対応できるのか、といった観点から評価を行うことができる。具体的には、これまでに起こった事故の記録などをもとに作った条件でのシミュレーションなどを繰り返すことにより、現状の自動運転の性能を評価することが可能となる。さらに、この評価を第三者が行うと共に、回避できそうな不測の事態についての情報を共有することで、自動運転に関する基本的な期待に関する社会的な合意形成を促進すると共に、例えば、事故が起こった場合の判断が、既存の事例と比較して、難しい判断だったのかどうか、といった評価が可能となる。

3.5 考察

本試論で提案した枠組は、一般道での車の自動運転のようにオープンな環境に製品を送り出す製造者とその利用者の責任分担を考えるための基準を示したものである。この枠組に基づいて責任分担を考えることで、社会通念上、存在を許容するリスクを考慮した責任者や利用者への責任の追求を行ったり、利用者がリスクを許容することによって、経済合理性の高い製品の利用を考えるといったことが可能となる。

本枠組には、リスクという、一般の利用者にとっては、あまり考えたくない情報の明示化を求めている点が難しいという考え方もある。しかし、これは、運転者がリスクに対して、無自覚的なだけであるだけであるとも考えられる。例えば、実際に車を運転する運転者には、各交差点で必要以上に周りを確認する人、完全に交通法規に従った運転をする人、道路の車の流れにあわせた運転をする人など様々なタイプの運転者が存在する。彼らの運転のスタイルをリスクという形に読替えるといった操作を行うことで、各運転者は、リスクに対して、より自覚的になるとともに、車への設定が可能になると考えている。この考え方は、製品が人工知能を搭載し、自律的に動く場合で

も、製品はあくまでも道具であり、道具を使う責任は、道具を使う人も分担するべきであるという立場に立つものである。

一方で、インターネット上で行われているトロッコ問題のような調査 [Bonnefon 16] は、主体的な判断を行う道具の責任を議論するものである。この議論は、直接的には、本試論で議論している枠組とは関係しないが、社会通念上、存在を許容するリスクや、そのリスクに応じた責任の取り方についての情報としては、非常に有用であると考えている。

4. まとめ

本試論では、オープンな環境で自律的に動く製品について、その製品が問題を起こしたときの責任の分担の枠組について議論を行ってきた。本枠組では、利用者や不測の事態に関係する他のステークホルダーが認識していたリスクについての情報を用いることにより、これらのステークホルダー間で責任を分担する。また、これまでの交通ルールなどを社会のルール社会通念上、存在を許容されるリスクとして、捉え直すことにより、リスクに関する議論に不馴れな利用者のために、リスクレベルについて考えることができる枠組を並行して提供することを考える。このような枠組を設定することで、不測の事態に対する責任分担の妥当性を、これまでの交通事故などでの過失割合などの議論を踏まえて判断することが可能となる。そのため、本枠組では、トロッコ問題などの議論で行われているような極端な議論ではなく、より柔軟な責任分担に関する議論を行う基盤として利用できるのではないか、と考えている。

参考文献

[Bonnefon 16] Bonnefon, J.-F., Shariff, A., and Rahwan, I.: The social dilemma of autonomous vehicles, *Science*, Vol. 352, No. 6293, pp. 1573–1576 (2016)

[Foot 67] Foot, P.: The Problem of Abortion and the Doctrine of Double Effect, *Oxford Review*, Vol. 5, pp. 5–15 (1967)