参考１

国際的な議論のための
ＡＩ開発ガイドライン案

平成 29 年 7 月 28 日

ＡＩネットワーク社会推進会議

目次

序文

ＡＩに関する研究開発や利活用は今後飛躍的に発展することが期待されている。こうした中、２０１６年４月に日本で開催されたＧ７情報通信大臣会合において、ホスト国である日本はＡＩ開発原則のたたき台を紹介し、各国関係閣僚による議論が行われた。その結果、Ｇ７において「ＡＩ開発原則」及びその内容の解説からなる「ＡＩ開発ガイドライン」の策定に向け、引き続きＧ７各国が中心となり、ＯＥＣＤ等国際機関の協力も得て議論していくことで合意した。

少子高齢化などの課題を抱える日本は、ＡＩを積極的に開発し利活用することにより様々な課題を解決するとともに、その知見を活かしつつＡＩの開発について留意することが期待される事項等を国際的に発信することにより、国際社会に大きく貢献することができる。

本ガイドライン案は、ＡＩの便益の増進及びリスクの抑制のため研究開発において留意することが期待される事項に関するＧ７やＯＥＣＤにおける国際的な議論のための基礎となる文書として作成されたものである。ＡＩに関する技術がその発展の途上にあることに鑑みれば、国際的な「ＡＩ開発原則」及びその内容の解説からなる「ＡＩ開発ガイドライン」は、規制の導入を目指すものとすることは適当ではない。そこで、本ガイドライン案は、非規制的で非拘束的なソフトローとして国際的に共有される指針の案として作成されたものである。こうした指針に盛り込むことが適当な内容を議論することを通じ、

- 「ＡＩ開発ガイドライン」及び「ＡＩ利活用ガイドライン」の策定に向け、ＡＩに関する研究開発や利活用に関係する多様なステークホルダ（開発者、サービス提供者、市民社会を含む利用者、各国政府、国際機関など）を含む国内及び国際的な議論を加速化すること
- ＡＩの研究開発及び利活用におけるベストプラクティスの国際的な共有を促すことにより、ＡＩへの利用者や社会の信頼を獲得し、ＡＩの研究開発及び利活用の円滑化に資すること

が期待される。

ＡＩ開発ガイドライン案[1]

１．　目的

ＡＩの研究開発や利活用は、今後急速に進展することが期待されているところであり、ＡＩネットワーク化（ＡＩシステムがインターネットその他の情報通信ネットワークと接続され、ＡＩシステム相互間又はＡＩシステムと他の種類のシステムとの間のネットワーク（以下において「ＡＩネットワーク」という場合がある 。）が形成されるようになることをいう。以下同じ。）が進展していく過程で、個人、地域社会、各国、国際社会[2]の抱える様々な課題の解決に大きく貢献するなど、人間及びその社会や経済に多大な便益を広範にもたらすことが期待される。このような方向に向けて、ＡＩの研究開発や利活用を加速化していくことが求められる。

その一環として、ＡＩシステムが社会や経済にもたらす便益の増進を図るとともに、不透明化や制御喪失などＡＩシステムに関するリスクの抑制を図る観点から、関連する社会的・経済的・倫理的・法的な課題に対応することが必要となる。特に、ＡＩシステムを利活用するサービスは、他の情報通信サービス同様、ネットワークを通じて国境を越えて提供されるものであることから、オープンな議論を通じ、国際的なコンセンサスを醸成し、非規制的で非拘束的なソフトローたるガイドラインやそのベストプラクティスをステークホルダ（開発者、サービス提供者、市民社会を含む利用者、各国政府、国際機関など）間で国際的に共有することにより、ＡＩシステムの便益の増進とリスクの抑制を図ることが求められる。

以上の問題意識に鑑み、本ガイドラインは、ＡＩネットワーク化の健全な進展を通じてＡＩシステムの便益の増進とリスクの抑制を図ることにより、利用者の利益を保護するとともにリスクの波及を抑止し、人間中心の智連社会[3]を

---

1 本ガイドライン案は、国際的に共有される「ＡＩ開発ガイドライン」の策定に向けた国際的な議論に用いられることを念頭に作成されたものである。

2 国際社会の抱える課題については、国連の「持続可能な開発目標」（SDGs）（http://www.un.org/ga/search/view_doc.asp?symbol=A/70/L.1）などを参照。

3 智連社会（Wisdom Network Society）とは、ＡＩネットワーク化の進展の結果として、人間がＡＩネットワークと共生し、データ・情報・知識を自由かつ安全に創造・流

実現することを目的とする。

そこで、本ガイドラインは、上記の目的を達成する観点から、今後のＡＩシステムの開発において留意することが期待される事項であるＡＩ開発原則及びその内容の解説を取りまとめたものである。

なお、ＡＩシステムの研究開発は多様な利活用の分野に及ぶものであり、分野ごとにＡＩシステムのもたらす便益やリスクは異なる可能性がある。このため、本ガイドラインでは、ＡＩの開発に関し留意することが期待される事項のうち、利活用の分野に共通する事項又は分野間の連携に関し留意することが期待される事項について定めることとし、分野ごとの事情に応じて留意することが期待される事項については、本ガイドラインとは別に、各々の分野ごとに当該分野に係るガイドラインの在り方に関し、策定の要否を含め、各分野の国際機関を含む関係するステークホルダによる議論が行われることが期待される。

また、ＡＩシステムは利活用の過程を通じて学習等により出力やプログラムが継続的に変化する可能性があることから、開発者が留意することが期待される事項のみならず、利用者が留意することが期待される事項も想定される。このため、本ガイドラインとは別にＡＩ利活用ガイドラインを策定することについて、国際的な議論を進めていくことが期待される。

２．　基本理念

本ガイドラインの目的に鑑み、次に掲げる理念を一体的なものとして本ガイドラインの基本理念とする。

１．人間がＡＩネットワークと共生することにより、その恵沢がすべての人によってあまねく享受され、人間の尊厳と個人の自律が尊重される**人間中心の社会を実現**すること。
２．ＡＩの研究開発と利活用が今後急速に発展し、ネットワーク化されたＡＩシステムが国境を越えて人間及び社会に広範かつ多大な影響を及ぼすものと見込まれることから、ＡＩシステムの研究開発の在り方について、非拘束的なソフトローたる**指針やそのベストプラクティスをステークホルダ間で国際**

通・連結して「智のネットワーク」（Wisdom Network）を構築することにより、あらゆる分野におけるヒト・モノ・コト相互間の空間を越えた協調が進展し、もって創造的かつ活力ある発展が可能となる社会である（【参考】「ＡＩネットワーク化と智連社会」参照）。

的に共有すること。

３．イノベーティブでオープンな研究開発と公正な競争を通じ、ＡＩネットワークの便益を増進するとともに、学問の自由や表現の自由といった民主主義社会の価値を最大限尊重しつつ、ＡＩネットワークにより権利利益が侵害されるリスクを抑制するため、**便益とリスクの適正なバランスを確保**すること。

４．ＡＩ関連の技術が引き続き急速に進展していくことが見込まれる中、**技術的中立性を確保**する観点から特定の技術や手法に基づくＡＩの研究開発を阻害しないよう配慮するとともに、**開発者にとって過度の負担とならないものとするよう留意**すること。

５．ＡＩ関連技術やＡＩの利活用が今後とも飛躍的に発展することが期待されることから、ＡＩネットワーク化の進展等を踏まえ、国際的な議論を通じて、**本ガイドラインを不断に見直し、必要に応じて柔軟に改定**すること。また、本ガイドラインの見直しに際しては関係するステークホルダの参画を得るなど、広範で柔軟な議論に努めること。

３．　用語の定義及び対象範囲

３－１　用語の定義

２．基本理念に鑑み、本ガイドラインにおける「ＡＩ」については、以下のとおり定義する。

「**ＡＩ**」とは、「ＡＩソフト及びＡＩシステムを総称する概念」をいう[4]。

●「**ＡＩソフト**」とは、データ・情報・知識の学習等[5]により、利活用の過程を通じて自らの出力やプログラムを変化させる機能を有するソフトウェ

---

[4] 本ガイドラインにおけるＡＩの定義は、現在既に実用化されている特化型ＡＩを主たる対象として想定しているが、自律性を有するＡＩや汎用ＡＩ（Artificial General Intelligence）の開発など今後予想されるＡＩに関する急速な技術発展を見据え、今後開発される多種多様なＡＩについても、学習等により自らの出力やプログラムを変化させる機能を有するものである場合には、含み得るものとしている。

本ガイドラインにおいては、基本理念４.に掲げる技術的中立性の確保の見地などから、上述のとおりＡＩを定義しており、今後開発される多種多様なＡＩについてもその機能次第で含み得るものとしている。なお、本ガイドラインにおけるＡＩの定義の在り方については、ＡＩの技術発展の動向等を踏まえ、今後継続的に議論を行っていくことが必要である。

[5] 学習以外の方法によりＡＩソフトが自らの出力やプログラムを変化させる要因としては、例えば、データ・情報・知識に基づく推論や、センサやアクチュエータ等を通じた環境とのインタラクションなどが考えられる。

アをいう。例えば、機械学習ソフトウェアはこれに含まれる。

●「**ＡＩシステム**」とは、ＡＩソフトを構成要素として含むシステムをいう。例えば、ＡＩソフトを実装したロボットやクラウドシステムはこれに含まれる。

ＡＩシステムの「開発者」及び「利用者」については、以下のとおり定義する。ただし、「開発者」及び「利用者」は場面に応じて個別に決まる相対的な概念であることに留意する必要がある。

●「**開発者**」とは、ＡＩシステムの研究開発（ＡＩシステムを利用しながら行う研究開発を含む。）を行う者（自らが開発したＡＩシステムを用いてＡＩネットワークサービスを他者に提供するプロバイダを含む。）をいう。

●「**利用者**」とは、ＡＩシステムを利用する者（最終利用者（エンドユーザ）のほか、他者が開発したＡＩネットワークサービスを第三者に提供するプロバイダを含む。）をいう。

３－２　対象範囲

本ガイドラインの対象とする**ＡＩシステムの範囲**は、ＡＩシステムがネットワークを通じて国境を越えて利用され、広く人間及び社会に便益やリスクをもたらす可能性があることから、ネットワーク化され得るＡＩシステム（ネットワークに接続可能なＡＩシステム）とする。

本ガイドラインの対象とする**開発者の範囲**は、本ガイドラインが非拘束的なソフトローとしての指針であることに鑑み、３－１で定義された開発者すべてとする。

本ガイドラインの対象とする**開発の範囲**は、学問の自由の尊重、社会に与える影響の大きさ等に鑑み、閉鎖された空間（実験室、セキュリティが十分に確保されたサンドボックス等）内での開発は対象とせず、ネットワークに接続して行う段階とする。

４．ＡＩ開発原則

（主にＡＩネットワーク化の健全な進展及びＡＩシステムの便益の増進に関する原則）

① **連携の原則**-----開発者は、ＡＩシステムの相互接続性と相互運用性に留意する。

（主にＡＩシステムのリスクの抑制に関する原則）

② **透明性の原則**-----開発者は、ＡＩシステムの入出力の検証可能性及び判断結果の説明可能性に留意する。

③ **制御可能性の原則**-----開発者は、ＡＩシステムの制御可能性に留意する。

④ **安全の原則**-----開発者は、ＡＩシステムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。

⑤ **セキュリティの原則**-----開発者は、ＡＩシステムのセキュリティに留意する。

⑥ **プライバシーの原則**-----開発者は、ＡＩシステムにより利用者及び第三者のプライバシーが侵害されないよう配慮する。

⑦ **倫理の原則**-----開発者は、ＡＩシステムの開発において、人間の尊厳と個人の自律を尊重する。

（主に利用者等の受容性の向上に関する原則）

⑧ **利用者支援の原則**-----開発者は、ＡＩシステムが利用者を支援し、利用者に選択の機会を適切に提供することが可能となるよう配慮する。

⑨ **アカウンタビリティの原則**------開発者は、利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める。

５．開発原則の解説

① 連携の原則-----開発者は、ＡＩシステムの相互接続性と相互運用性に留意する。

（解説）

開発者は、ＡＩ ネットワーク化の健全な進展を通じて、ＡＩ システムの便益を増進するとともに、リスクの抑制に関する複数の開発者の取組が相互に調和して有効に機能するよう、ＡＩシステムの多様性を踏まえつつ、自らの開発するＡＩシステムと他のＡＩシステム等との相互接続性と相互運用性[6]に留意

[6] ここで相互運用性と相互接続性としては、自らの開発するＡＩシステムが情報通信ネッ

することが望ましい。そのため、開発者は、以下の事項について留意することが望ましい。

- ●相互接続性と相互運用性を確保するために有効な関連情報の共有に向けて協力するよう努めること。
- ●国際的な標準や規格がある場合には、これに準拠したＡＩシステムを開発するよう努めること。
- ●データ形式の標準化、ＡＰＩを含むインターフェースやプロトコルのオープン化に対応するよう努めること。
- ●自らの開発するＡＩシステムと他のＡＩシステム等との相互接続・相互運用により意図しない事象が生ずるリスクに留意すること。
- ●ＡＩの開発に関連する知的財産に関し、保護と利活用のバランスに配慮しつつ、標準必須特許などＡＩシステムと他のＡＩシステム等との相互接続性・相互運用性の確保に資する知的財産権のライセンス契約及びその条件についてオープンかつ公平な取扱いを図るよう努めること。

②透明性の原則――――開発者は、ＡＩシステムの入出力の検証可能性及び判断結果の説明可能性に留意する[7]。

（解説）

本原則の対象となるＡＩシステムとしては、利用者及び第三者の生命、身体、自由、プライバシー、財産などに影響を及ぼす可能性のあるＡＩシステムが想定される。

開発者は、ＡＩシステムに対する利用者を含む社会の理解と信頼が得られるよう、採用する技術の特性や用途に照らし合理的な範囲で、ＡＩシステムの入出力の検証可能性及び判断結果の説明可能性に留意することが望ましい。

③制御可能性の原則――――開発者は、ＡＩシステムの制御可能性に留意する。

（解説）

開発者は、ＡＩシステムの制御可能性に関するリスクを評価するため、あら

トワークと接続され、他のＡＩシステム等と相互に適切に協調して運用することが可能であることを念頭に置いている。

[7] 本原則は、開発者によるアルゴリズム、ソースコード、学習データの開示を想定するものではない。また、本原則の解釈に当たっては、プライバシーや営業秘密への配慮も求められる。

かじめ検証及び妥当性の確認[8]を行うよう努めることが望ましい[9]。こうしたリスク評価の手法としては、社会において実用化される前の段階において、実験室内やセキュリティが確保されたサンドボックスなどの閉鎖空間において実験を行うことが考えられる。

また、開発者は、制御可能性を確保するため、採用する技術の特性に照らして可能な範囲において、人間や信頼できる他のＡＩによる監督（監視、警告など）や対処（ＡＩシステムの停止、ネットワークからの切断、修理など）の実効性に留意することが望ましい。

④安全の原則――開発者は、ＡＩシステムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。

（解説）

本原則の対象となるＡＩシステムとしては、アクチュエータ等を通じて利用者及び第三者の生命、身体、財産に危害を及ぼす可能性のあるＡＩシステムが想定される。

開発者は、本原則に関する国際標準等を参照するとともに、特にＡＩシステムが学習等によって出力やプログラムが変化する可能性を踏まえ、以下の事項について留意することが望ましい。

- ●ＡＩシステムの安全性に関するリスクを評価・抑制するため、あらかじめ検証及び妥当性の確認を行うよう努めること。
- ●ＡＩシステムがアクチュエータ等を通じて稼動する際の本質安全（アクチュエータの運動エネルギーなど本質的な危険要因の低減）や機能安全（自動ブレーキなど付加的な制御装置の作動によるリスクの抑制）に資するよう、ＡＩシステムの開発の過程を通じて、採用する技術の特性に照らし可能な範囲で措置を講ずるよう努めること。
- ●ＡＩシステムを利用する際の利用者及び第三者の生命、身体、財産の安全に関する判断（例えば、ＡＩを搭載したロボットの事故発生時に、優先的

---

[8] 検証（verification）及び妥当性の確認（validation）は、あらかじめリスクを評価し抑制するための手法であるが、 前者は形式的な整合性の確認を意味して用いられるのに対し、後者は実質的な妥当性の確認を意味して用いられることが一般的である（*See e.g.*, The Future of Life Institute（ＦＬＩ）, *Research Priorities for Robust and Beneficial Artificial Intelligence* (2015)。

[9] リスク評価の要素としては、例えば、ＡＩシステムが与えられた目標を形式的に達成するために開発者の意図に実質的に反する動作（報酬ハッキング）を行うリスクやＡＩシステムが学習等による利活用の過程を通じた変化に伴い開発者の意図しない動作を行うリスク等に配慮することが考えられる。*See e.g.*, Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety*, arXiv:1606.06565 [cs.AI] (2016).

に保護される生命、身体、財産の順位などに関する判断）を行うＡＩシステムを開発する場合において、利用者等ステークホルダに対して当該ＡＩシステムの設計の趣旨及びその理由を説明するよう努めること。

⑤セキュリティの原則-----開発者は、ＡＩシステムのセキュリティに留意する。

（解説）

開発者は、ＯＥＣＤセキュリティガイドラインなどセキュリティに関する国際的な指針を踏まえるほか、ＡＩシステムが学習等によって出力やプログラムが変化する可能性があることを踏まえ、以下の事項について留意することが望ましい。

- ●ＡＩシステムの情報セキュリティについては、通常、情報の機密性、完全性及び可用性が確保されることが求められるが、必要に応じて、ＡＩシステムの信頼性（意図したとおりに動作が行われ、権限を有しない第三者による操作を受けないこと）や頑健性（物理的な攻撃や事故への耐性）にも留意すること。
- ●ＡＩシステムのセキュリティに関するリスクを評価・抑制するため、あらかじめ検証や妥当性確認を行うよう努めること。
- ●ＡＩシステムの開発の過程を通じて、採用する技術の特性に照らし可能な範囲でセキュリティ対策を講ずるよう努めること（セキュリティ・バイ・デザイン）。

⑥プライバシーの原則-----開発者は、ＡＩシステムにより利用者及び第三者のプライバシーが侵害されないよう配慮する。

（解説）

本原則にいうプライバシーの範囲には、空間に係るプライバシー（私生活の平穏）、情報に係るプライバシー（個人データ）及び通信の秘密が含まれる。開発者は、ＯＥＣＤプライバシーガイドラインなどプライバシーに関する国際的な指針を踏まえるとともに、ＡＩシステムが学習等によって出力やプログラムが変化する可能性があることを踏まえ、以下の事項について留意することが望ましい。

- ●プライバシー侵害のリスクを評価するため、あらかじめプライバシー影響評価を行うよう努めること。
- ●ＡＩシステムの利活用時におけるプライバシー侵害を回避するため、当該システムの開発の過程を通じて、採用する技術の特性に照らし可能な範囲で措置を講ずるよう努めること（プライバシー・バイ・デザイン）。

⑦倫理の原則――開発者は、ＡＩシステムの開発において、人間の尊厳と個人の自律を尊重する。

（解説）

開発者は、人間の尊厳と個人の自律を尊重するに当たり、人間の脳や身体と連携するＡＩシステムを開発する場合は、生命倫理に関する議論などを参照しつつ、特に慎重に配慮することが望ましい。

開発者は、採用する技術の特性に照らし可能な範囲で、ＡＩシステムの学習データに含まれる偏見などに起因して不当な差別が生じないよう所要の措置を講ずるよう努めることが望ましい。

開発者は、国際人権法や国際人道法を踏まえ、ＡＩシステムが人間性の価値を不当に毀損することがないよう留意することが望ましい。

⑧利用者支援の原則――開発者は、ＡＩシステムが利用者を支援し、利用者に選択の機会を適切に提供することが可能となるよう配慮する。

（解説）

開発者は、ＡＩシステムの利用者を支援するため、以下の事項について留意することが望ましい。

- ●利用者の判断に資する情報を適時適切に提供し、かつ、利用者にとって操作しやすいインターフェースが利用可能であることに配慮するよう努めること。
- ●利用者に選択の機会を適時適切に提供する機能（例えば、デフォルトの設定、理解しやすい選択肢の提示、フィードバックの提供、緊急時の警告、エラーへの対処など）が利用可能であることに配慮するよう努めること。
- ●ユニバーサルデザインなど社会的弱者の利用を容易にするための取組に努めること。

また、開発者は、ＡＩシステムの学習等による出力又はプログラムの変化の可能性を踏まえ、利用者に対し適切な情報提供を行うよう努めることが望ましい。

⑨アカウンタビリティの原則――開発者は、利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める。

（解説）

開発者は、ＡＩシステムへの利用者や社会の信頼を得られるよう、自らの開発するＡＩシステムについてアカウンタビリティを果たすことが期待される。

具体的には、利用者にＡＩシステムの選択及び利活用に資する情報を提供す

るとともに、利用者を含む社会によるＡＩシステムの受容性を向上するため、開発者は、本ガイドラインの掲げる開発原則①～⑧の趣旨に鑑み、利用者等に対し自らの開発するＡＩシステムの技術的特性について情報提供と説明を行うほか、多様なステークホルダとの対話を通じて様々な意見を聴取するなど、ステークホルダの積極的な関与（フィードバック）を得るよう努めることが望ましい。

また、開発者は、自らの開発するＡＩシステムによってサービスを提供するプロバイダ等と情報を共有し、協力するよう努めることが望ましい。

【別添】関係するステークホルダに期待される役割

本ガイドラインの目的に鑑み、関係する産学民官のステークホルダは、例えば以下のような役割を果たすことが期待される。

１．各国政府及び国際機関は、本ガイドラインの運用や見直しにおいて各国政府、国際機関、開発者、市民社会を含む利用者など**多様なステークホルダ間の対話の促進**に向けた環境整備に努めることが期待される。
２．開発者、市民社会を含む利用者など関係するステークホルダは、上記１の対話に参加するとともに、本ガイドラインに適う**ベストプラクティスを共有**し、ＡＩをめぐる議論の多様性を確保しつつ、ＡＩの便益の増進及びリスクの抑制について、認識の共有を図り、相互に協力するよう努めることが期待される。
３．標準化団体等は、本ガイドラインに適う**推奨モデルを作成し公表**することが期待される。
４．各国政府は、**ＡＩの開発者コミュニティを支援**し、本ガイドラインに掲げるＡＩのもたらす便益の増進やリスクの抑制といった課題の解決に向けた取組と併せ、**ＡＩに関する研究開発を支援する政策**を積極的に推進することが期待される。

（参考）

AI ネットワーク化と智連社会

AIネットワーク化

1．AIシステムが、他のAIシステムとは連携せずに、インターネットその他の情報通信ネットワークを介して単独で機能。

2．複数のAIシステム相互間のネットワークが形成され、ネットワーク上のAIシステムが相互に連携して協調。

3．センサやアクチュエータを構成要素として含むAIネットワークが人間の身体又は脳と連携することを通じて、人間の潜在的能力が拡張。

4．人間とAIネットワークが共生し、人間社会のあらゆる場面においてシームレスに連携。

智連社会

智連社会（Wisdom Network Society 【WINS】）は、人間がAIネットワークと共生し、**データ・情報・知識を自由かつ安全に創造・流通・連結して「智のネットワーク」（Wisdom Network）を形成**することにより、**あらゆる分野におけるヒト・モノ・コト相互間の空間を越えた協調が進展**し、もって創造的かつ活力ある発展が可能となる**人間中心の社会像**。

人間がAIネットワークと共生し

データ・情報・知識を自由かつ安全に創造・流通・連結して「智のネットワーク」を形成することにより

あらゆる分野におけるヒト・モノ・コト相互間の空間を越えた協調が進展し

人機共生

総智連環

協調遍在

創造的かつ活力ある発展が可能となる社会