別紙２

「ＡＩ開発ガイドライン」（仮称）の策定に向けた国際的議論の用に供する素案の作成に関する論点（要旨）

平 成 ２ ８ 年 １ ２ 月 ２ ８ 日

ＡＩネットワーク社会推進会議事務局

（総務省情報通信政策研究所調査研究部）

論点の要旨（1／8）

第一　基本概念の定義

○「開発者」とは、AIの研究開発（AIを研究し、又は開発する行為のほか、複数のAIを組み合わせて一体的なAIとして機能するよう構成する行為を含む。この資料において、「AIの研究開発」を「AIの開発」又は単に「開発」という場合がある。）をする者（自らが研究開発したAIを実装するAIネットワークシステムによるAIネットワークサービスのプロバイダを含む。）をいうものとしてはどうか。

○「利用者」とは、他の開発者が開発したAI（他の開発者が開発したAIを実装するAIネットワークシステムにより他のプロバイダが提供するAIネットワークサービスを含む。）の提供を受けてAIネットワークシステムを利用する者（自らが構築するAIネットワークシステムを自ら利用する個人又は団体（最終利用者のほか、AIネットワークサービスを他の者に提供するプロバイダを含む。）のほか、プロバイダからAIネットワークサービスの提供を受ける最終利用者をも含む。）をいうものとしてはどうか。

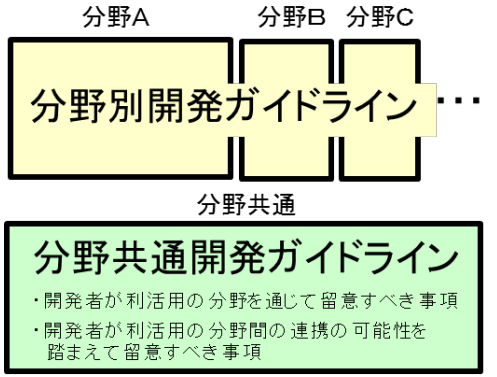
○ただし、「開発者」及び「利用者」は、関係者間の関係に即して、関係が生ずる場面ごとに個別に評価される相対的な概念。

第二　ＡＩ開発ガイドラインの体系

○開発ガイドラインの体系は、「分野共通開発ガイドライン」及び「分野別開発ガイドライン」からなるものとしてはどうか。

○「分野共通開発ガイドライン」は、本推進会議が検討と議論を推進することとしてはどうか。

○「分野別開発ガイドライン」は、その策定の要否も含め、分野ごとの国際機関を含む関係ステークホルダー自身による検討と議論に委ねることとしてはどうか。



第三　分野共通開発ガイドラインの構成

○分野共通開発ガイドライン　（OECDのガイドラインであれば、理事会勧告の附属文書（Annex））

AIネットワークシステムの構成要素となり得るAIの開発者がその研究開発に当たり、AIネットワーク化の健全な進展の促進並びにAIネットワークシステムの便益の増進及びリスクの抑制に関し、AIネットワークシステムの利活用の分野を通じて留意すべき事項及び利活用の分野間の連携の可能性を踏まえて留意すべき事項に関する原則（「開発原則」）並びにその説明

分野共通ガイドラインの構成（たたき台）

1. 目的
2. 基本理念（智連社会の構成要素、智連社会の基本理念等）
3. 定義（AI、AIネットワークシステム、AIネットワークサービス、開発者、利用者）
4. 適用範囲（開発者、AI、研究開発の段階）
5. 原則 → ○○の原則　××を△△すべきである。【説明】・・・・・・・・である。

○分野共通開発ガイドラインの関連文書　（OECDのガイドラインであれば、理事会勧告の本紙）

・分野共通開発ガイドラインに定める事項に関連し、国、関係国際機関等に推奨すべき事項

・ガイドラインの見直しの時期及び方法

論点の要旨（2／8）

第四　分野共通開発ガイドラインの目的、基本理念等

○分野共通開発ガイドラインの「目的」として、次に掲げる趣旨を、必要に応じて再構成した上で掲げることとしてはどうか。

> このガイドラインは、AIネットワークシステムの公共性に鑑み、AIネットワークシステムの構成要素となり得るAIの研究開発を行う者が、その研究開発に当たり、AIネットワーク化の健全な進展の促進並びにAIネットワークシステム（AIネットワークサービスを含む。以下同じ。）の便益の増進及びリスクの抑制に関し、AIネットワークシステムの利活用の分野を通じて又は分野間の連携の可能性を踏まえて留意すべき事項を開発原則として整理し、非拘束的な枠組みとして国際的に共有することにより、AIネットワークシステムの最終利用者の利益を保護するとともに第三者及び社会への波及的な悪影響を防止し、もって人間中心の智連社会の形成に資することを目的とする。

○分野共通開発ガイドラインの策定及び解釈に当たっての「基本理念」として、①智連社会の構成要素、②智連社会の基本理念、③リスクへの適時適切な対処、④関係する価値・利益のバランスの確保及び⑤AIネットワーク化の進展及び関連するリスクの顕在化に応じた開発原則・開発ガイドラインの見直しの５点を、必要に応じて再構成した上で分野共通開発ガイドラインに掲げることとしてはどうか。

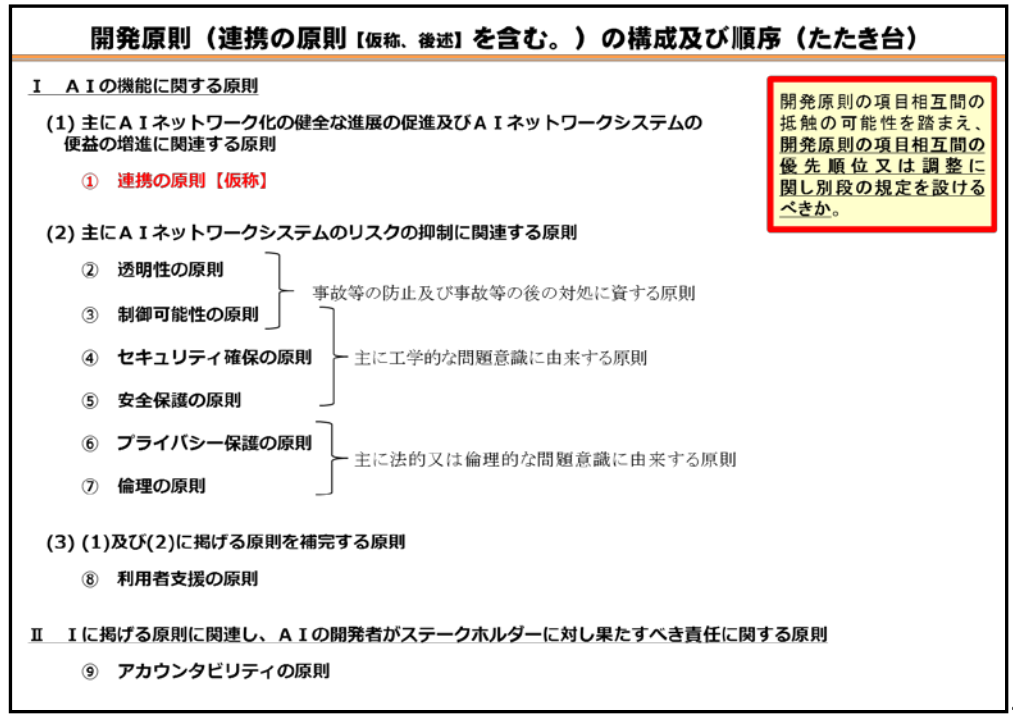
第五　分野共通開発ガイドラインの適用範囲

○適用対象とする「開発者」の範囲は、限定する必要はないのではないか。

○適用対象とする「AI」の範囲は、その機能如何にかかわらず、AIネットワークシステムの構成要素となり得るAI、すなわち、何らかの情報通信ネットワークシステムに実装し又は接続し得るAIを広く包含することとしてはどうか。

○適用対象とする「研究開発の段階」については、学問の自由等に鑑み、閉鎖された空間（実験室等）の外につながる情報通信ネットワークシステムに実装し又は接続して行う段階に限定することとしてはどうか。

第六　開発原則の構成及び順序



論点の要旨（3／8）

第七　開発原則の個々の項目の内容の具体化

(1) 透明性の原則　ＡＩネットワークシステムの動作の検証可能性及び説明可能性を確保すること。

○**動作の透明性**（検証可能性及び説明可能性）**が要請されるAIの動作の範囲**如何。**入出力、通信及び判断**としてよいか。

○**個人の生命・身体の安全等重要な権利利益若しくは法益に関するリスクを惹起し得る、又は個人に関する重大な決定のために利活用されるAIネットワークシステムの構成要素となり得るAIに関し、**

- 当該AIの動作のうち、**入出力及び通信については検証可能性を確保すべき**とし、**判断（推論メカニズム、データの学習履歴等）については、深層学習等における判断過程のブラックボックス化が指摘されていることなどに鑑み、技術的及び経済的な事情に鑑み合理的な範囲・水準で、検証可能性を確保するよう努めるべき**としてはどうか。
- **技術的及び経済的な事情に鑑み合理的な範囲・水準で動作の説明可能性を確保するよう努めるべき**としてはどうか。

(2) 制御可能性の原則　ＡＩネットワークシステムの制御可能性を確保すること。

○**制御不能となるリスクにつき、その蓋然性が高い又は不確実と考えられるAIについては、一般社会で利用される前に、実験室等閉鎖された空間において、当該空間の外につながる情報通信ネットワークシステムに接続せずに、AIの制御可能性について実験を行い、リスク評価を行うことにより、制御可能性を確保すべき**としてはどうか。

○AIネットワークシステムの制御可能性を継続的に確保するために、その構成要素となり得る**AIについて、人間又は信頼し得る他のAIによる監督及び対処（停止、切断、修理等）の実効性を確保すべき**としてはどうか。

○【上記二点のほか、次頁の（※）において、「(3) セキュリティ確保の原則」及び「(4) 安全保護の原則」と共通の論点を所掲。】

(3) セキュリティ確保の原則　ＡＩネットワークシステムの頑健性及び信頼性を確保すること。

○**セキュリティの範囲**には、**情報の機密性、完全性及び可用性のみならず**、当該システムの**信頼性**（意図した通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）**及び頑健性**（物理的な攻撃や事故への耐性）**の維持も含まれる**としてはどうか。

○AIネットワークシステムのセキュリティを確保することができるよう、その構成要素となり得る**AIの設計段階において措置（セキュリティ・バイ・デザイン）を講ずべき**としてはどうか。

○【上記二点のほか、次頁の（※）において、「(2) 制御可能性の原則」及び「(4) 安全保護の原則」と共通の論点を所掲。】

論点の要旨（4／8）

(4) 安全保護の原則　ＡＩネットワークシステムが利用者及び第三者の生命・身体の安全に危害を及ぼさないように配慮すること。

○**安全保護の原則が適用されるAIの範囲は、個人の生命・身体の安全に関するリスクを惹起し得るAIネットワークシステムの構成要素となり得るAI**としてはどうか。

○AIネットワークシステムにおける**本質安全（運動能力等の抑制）、制御安全（監視装置等の実装）、機能安全等を確保することができるよう**、その構成要素となり得る**AIの設計段階において措置（セーフティ・バイ・デザイン）を講ずべき**としてはどうか。

○**ＡＩネットワークシステムを利活用する際の利用者及び第三者の生命・身体の安全に関する判断（例：生命・身体の安全を保護される個人の優先順位等に関する判断）を行うAIを研究開発する場合には、開発者は利用者等に対し当該判断を行うAIに関する設計の趣旨及び理由を説明すべき**としてはどうか。

○【上記三点のほか、下記（※）において、「(2) 制御可能性の原則」及び「(3) セキュリティ確保の原則」と共通の論点を所掲。】

（※） 上記「(2) 制御可能性の原則」、「(3) セキュリティ確保の原則」及び「(4) 安全保護の原則」の共通の論点

○**「(2) 制御可能性の原則」、「(3) セキュリティ確保の原則」及び「(4)安全保護の原則」においては**、リスクを評価し抑制するため、AIネットワークシステムの構成要素となり得るAIについて、**予め制御可能性の検証**（verification）[※形式的な整合性の検証]**及び妥当性確認**（validation）[※実質的な妥当性の確認]**を行うことが必要となる旨を定めるべき**ではないか。

(5) プライバシー保護の原則　ＡＩネットワークシステムが利用者及び第三者のプライバシーを侵害しないように配慮すること。

○プライバシー保護の原則において配慮されるべき**プライバシーの範囲には、空間プライバシー**（私生活の平穏）、**情報プライバシー**（個人データ）、**通信の秘密及び生体プライバシーが含まれる**としてはどうか。

○AIネットワークシステムにおけるプライバシー侵害のリスクを評価するために、その構成要素となり得るAIについて**予めプライバシー影響評価を行うべき**としてはどうか。

○AIネットワークシステムがその利活用に当たりプライバシーが保護されるものとなるよう、その構成要素となり得る**AIの設計段階において措置（プライバシー・バイ・デザイン）を講ずべき**としてはどうか。

論点の要旨（5／8）

(6) 倫理の原則　ＡＩネットワークシステムの研究開発において、人間の尊厳と個人の自律を尊重すること。

○倫理の原則においては、**人間性（humanity）の価値を中心に据えつつ、人間の尊厳と個人の自律を尊重すべき**旨を掲げることとしてはどうか。

○**国際人権法・国際人道法等を参照し、AIネットワークシステムが人間性の価値を毀損してはならない**としてはどうか。

○**人間の脳・身体と融合又は連携するAIを研究開発する際には、人間の尊厳と個人の自律の尊重について、生命倫理等の議論も参照しつつ、特に慎重に配慮すべき**としてはどうか。

○AIの開発において、**個人を公平に尊重**する観点から、**技術的に可能な範囲で、AIの学習するデータに含まれる偏見等に起因する差別**（人種、性、宗教等による差別）**を防止するための措置を講ずべき**としてはどうか。

(7) 利用者支援の原則　ＡＩネットワーク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よう配慮すること。

○（最終）利用者に操作されるAIネットワークシステムの構成要素となり得るAIについては、

- **利用者に対し適時適切にその判断に資する情報を提供し、かつ、利用者にとって操作しやすいインターフェイスが利用可能となるよう設計すべき**ではないか。
- **利用者に選択の機会を適時適切に提供する機能（ナッジ：例えば、デフォルトの設定、理解しやすい選択肢の提示・体系化、フィードバックの提供、緊急時の警告、エラーへの対処等）が利用可能となるよう設計すべき**ではないか。
- **ユニバーサル・デザイン等社会的弱者の受容可能性を高めるための取組に努めるべき**としてはどうか。

(8) アカウンタビリティの原則　ＡＩネットワークシステムの研究開発者が利用者など関係するステークホルダーに対しアカウンタビリティを果たすこと。

○開発者が説明責任を果たす上では、**特に利用者等に対し開発原則の遵守状況等について説明を行うべきであるほか、多様なステークホルダーと対話を行ってその意見を聴取する等ステークホルダーの積極的な関与を得るべき**ではないか。

○**開発原則の遵守状況につき開発者から説明を受けた利用者によるAIネットワークシステムへの信頼・期待が保護されるよう、利用者の責任との関係に留意しつつ、開発者の責任の在り方について指針を示すべき**ではないか。

論点の要旨（6／8）

第八　連携の原則【仮称】

○AIの多様性を踏まえつつ、相互接続性・相互運用性の確保等AIネットワークシステム相互間の円滑な連携の確保に関し開発者が留意すべき事項を「連携の原則」【仮称】として開発原則に追加することとしてはどうか。連携の原則【仮称】及びその説明において記すべき事項如何。

○AIネットワークシステム相互間の円滑な連携の確保に関し開発者が留意すべき事項に関連し、国、関係国際機関等に推奨すべき事項として分野共通開発ガイドラインの関連文書（OECDのガイドラインであれば、理事会勧告の本紙）に記すべき事項如何。

○上記二点に関連し、AIネットワークシステム相互間の連携がAIネットワークシステムの利活用に伴うものであることに鑑み、AIネットワークシステム相互間の円滑な連携の確保に関しAIネットワークシステムの利活用の段階において利用者（特にAIネットワークサービスのプロバイダ）が留意すべき事項及び国、関係国際機関等に推奨すべき事項を後述する分野共通利活用ガイドライン及びその関連文書（OECDのガイドラインであれば、理事会勧告の本紙）にそれぞれ記すこととしてはどうか。

なお、AI及びAIネットワークシステムが現時点においては揺籃期であることから、上記三点に関し法的規制の創設を検討することは、現時点では時期尚早ではないか。少なくとも関連する弊害の蓋然性が顕著になるまでは、分野共通開発ガイドライン及び分野共通利活用ガイドライン並びにそれぞれの関連文書により開発者及びAIネットワークサービスのプロバイダ等利用者が留意すべき事項並びに国、関係国際機関等に推奨すべき事項を国際的に共有した上で、国、関係国際機関等は、関連する動向を注視して、動向やベストプラクティスに関する情報を国際的に共有するとともに、AIネットワークシステム相互間の連携をめぐる紛争の発生状況等に応じて、国内の紛争及び国境を越えた紛争の処理の在り方等を検討して所要の措置を講ずるにとどめるような謙抑的な姿勢であるべきではないか。

第九　開発原則の実効性の確保の在り方

○分野共通開発ガイドラインの関連文書（OECDのガイドラインであれば、理事会勧告の本紙）において、国、関係国際機関等に対し、開発原則の実効性の確保のための方策として、例えば次に掲げる方策を検討するよう推奨する旨を記すこととしてはどうか。

（1）公共調達の対象とするAI及び公的研究費の交付対象とするAIに関し、開発原則を踏まえて条件を設定

（2）市場の機能を活用して、開発原則に適合しているAIが市場において利用者に選択されやすくなる環境を整備

（→(2)については、「第十　開発原則の実効性の確保における市場の活用の在り方」参照。）

○分野共通開発ガイドラインの関連文書（OECDのガイドラインであれば、理事会勧告の本紙）において、各国の関係機関、関係国際機関等に対し、開発原則の実効性に関する状況、実効性の確保に関するベストプラクティス等に関する情報を共有し、相互に協力するよう推奨する旨を記すこととしてはどうか。

論点の要旨（7／8）

第十　開発原則の実効性の確保のための市場の活用の在り方

○「第九　開発原則の実効性の在り方」の(2)に掲げる方策として、**分野共通開発ガイドラインの関連文書**（OECDのガイドラインであれば、理事会勧告の本紙）**において**、**各国に対し**、例えば**次の①及び②の仕組みを分野共通の一般的仕組みとして一体的に整備**することについて検討するよう**推奨**するとともに、**各分野に係る各国の関係機関、関係国際機関等に対し、利活用の分野ごとの事情に照らし、必要に応じ分野別の特則的仕組みを検討するよう推奨**する旨を記すこととしてはどうか。

① **開発者が**その開発するAIに関し**開発原則への適合性に関する情報を客観的に信頼できる形で自発的に提供する仕組み**

（例） 開発者が自発的に提供する情報に基づき、第三者機関が当該AIの開発原則への適合性を評価して認証する制度

② 上記**①の仕組みにより開発者が提供した情報において開発原則に適合しているとされているAI**を実装するAIネットワークシステムの利活用に伴い、当該AI**のリスクが顕在化したことに起因する第三者の被害等に関し、その利用者の法的責任、法的義務等が問題となる場合において**、**当該利用者の当該情報に対する信頼に基づく期待を保護**するための仕組み

（例）当該被害等に関する当該利用者の法的責任等を減免する制度

○開発原則の実効性を確保するために市場を活用する場合であっても、**開発原則に掲げる事項のうち、人権等他の利益とバーターにすべきでないものについては、AIネットワークシステムの用途に照らし、必要に応じ当該用途に係る利活用の分野に関連する制度の整備等を検討するよう当該分野に係る各国の関係機関、関係国際機関等に推奨**する旨を分野共通開発ガイドラインの関連文書（OECDのガイドラインであれば、理事会勧告の本紙）及び後述する分野共通利活用ガイドラインの関連文書（同前）に記すこととしてはどうか。

論点の要旨（8／8）

第十一　ＡＩネットワークシステムの利活用に関し利用者等が留意すべき事項

○AIネットワークシステムの利活用に関し利用者（AIネットワークサービスのプロバイダ、最終利用者）が留意すべき事項及び国、関係国際機関等に推奨すべき事項を整理して、国際的に共有する枠組みとして「AIネットワークシステム利活用ガイドライン」（仮称）及びその関連文書（OECDのガイドラインであれば、理事会勧告の本紙）を策定し、開発ガイドライン及びその関連文書と相互に補完し合う二本柱とすることに向け、OECD等の協力の下、国際的に議論すべきではないか。

○利活用ガイドラインの体系については、開発ガイドラインと同様に、分野共通ガイドライン及び分野別ガイドラインからなるものとすることが適当ではないか。

→以下両者を区別する場合には、前者を「分野共通利活用ガイドライン」といい、後者を「分野別利活用ガイドライン」という。

分野共通利活用ガイドラインは、AIネットワークシステム（AIネットワークサービスを含む。）の利用者（AIネットワークサービスのプロバイダ及び最終利用者を含む。）が、その利活用（AIネットワークサービスの提供及び利活用を含む。）に当たり、利活用の分野を通じて留意すべき事項及び分野間の連携の可能性を踏まえて留意すべき事項（「利活用原則」）並びにその説明を策定するものとして、本推進会議がその検討と議論を推進してはどうか。

分野別利活用ガイドラインは、各分野における策定の要否そのもの及び策定する場合における内容の双方ともに、各分野の関係国際機関を含む当該分野の産学民官のステークホルダー自身による検討と議論に委ねることとしてはどうか。

○分野共通利活用ガイドラインに定める「利活用原則」は、AIネットワークシステム（AIネットワークサービスを含む。）の利用者（AIネットワークサービスのプロバイダ及び最終利用者を含む。）が、その利活用（AIネットワークサービスの提供及び利活用を含む。）に当たり、次の①～③に掲げる見地から利活用の分野を通じて又は分野間の連携の可能性を踏まえて留意すべき事項としてはどうか。

① AIネットワーク化の健全な進展の促進及びAIネットワークシステムの便益の増進
② AIネットワークシステムのリスクの抑制
③ AIネットワークシステムの利活用に伴い、当該AIネットワークシステムに実装するAIのリスクの顕在化に起因する被害に関する被害者の利益の保護

○自らAIネットワークシステムを構築する最終利用者、自ら構築するAIネットワークシステムによりAIネットワークサービスを最終利用者等他の者に提供するプロバイダ及びプロバイダからAIネットワークサービスの提供を受ける最終利用者の種別に応じて、適用すべき利活用原則の範囲、内容等に異同があり得ることから、その異同を利活用ガイドラインに明記するとともに、これら利用者の種別ごとに整理したマニュアル等を作成することとしてはどうか。