別紙２－２

報告書２０１７（案）に関する意見募集に寄せられた主な意見[1]に対する考え方

番号	意見の要旨	意見に対する考え方（案）
「国際的な議論のためのＡＩ開発ガイドライン案」における用語の定義及び対象範囲について		
1	「ＡＩネットワーク化」の定義の明確化が必要である。また、定義に利用されている用語（具体的には、「データ」、「情報」、「知識」、「学習」）の定義の明確化が必要である。	「ＡＩネットワーク化」は、報告書本体３頁において「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が形成されるようになること」と定義しているが、御指摘を踏まえ、本ガイドライン案においても定義を記載することとした。また、「データ」、「情報」、「知識」、「学習」等の関係については、報告書本体４頁に記載している。
2	汎用ＡＩは、様々な技術的課題があり、開発・実用化の目処は立っておらず、実現されるとしてもかなり先のことと考えられている。遠い将来に開発されるものまで含むこととなれば、研究開発の自由を制約するおそれがある。 現時点で汎用ＡＩを含めることについては、必要性が乏しいものである上に、負の影響をもたらすおそれがあることから慎重の上にも慎重であるべきであり、技術開発の動向を見極めながら今後の議論に委ねるべきである。	本ガイドラインにおけるＡＩの定義は、現在すで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。 これは、以下に掲げる事項などに鑑みたものである。 ● 本ガイドラインにおけるＡＩの定義を行うに当たっては、特定の技術や手法に基づくＡＩのみを念頭に置くことや特定の技術や手法に基づくＡＩを除外することは、基本理念４に掲げる技術的中立性の確保の見地から適当ではないこと ● ＡＩの技術発展はスピードが早くその方向性も多様かつ不確実であると指摘されていること ● 専門家の間でも汎用ＡＩの実現時期については予測が分かれていること ● 特化型ＡＩと汎用ＡＩは連続的なものであるとの見方も有力であり、両者を峻別することは必ずしも容易ではないこと ● 自律性を有するＡＩや汎用ＡＩについては、将来的に実現した場合に、特化型ＡＩ以上に、社会や人間に広範なリスクを及ぼすおそれがあるのではないかとの懸念や不安も示されており、ＡＩへの利用者や社会の信頼を獲得するためには、このよ

[1] 提出された意見の全部について、次に掲げるＵＲＬのウェブサイトに所掲。
＜調整中＞

		うな懸念や不安に応接する必要があること ● 本ガイドラインは非拘束的な枠組みとして国際的に共有されることを目指していること ● 国内外におけるＡＩの開発の在り方に関する検討においては、特化型ＡＩのみならず、自律性を有するＡＩや汎用ＡＩも射程に入れて検討を行うものが有力となっていること その上で、御指摘や当該論点に関連する構成員の御意見等も踏まえ、「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」と加筆した。
３	「開発者」と「研究者」を明確に分けて議論すべきである。先端技術を研究しようとする研究者については、適用除外とした方がよい。その際には、「先端ＡＩ研究ガイドライン」等を別途策定すべきであると考える。	今日のＡＩの研究開発においては、研究と開発が一体的ないし連続的に行われていることも多く、開発者と研究者を峻別することは困難なのではないか。したがって、本ガイドラインとは別に、研究者のみを想定したガイドラインを策定することは適当ではないのではないか。 なお、本ガイドラインは、ＡＩの研究開発を拘束するものでなく、ＡＩの研究開発において開発者が留意することが期待される事項を整理したものである。
４	完全にネットワークから遮断した環境を構築することは、大学等の研究室レベルでは負荷が高く現実的ではない。ネットワークを利用した研究、実験を行うことが困難になる。被験者など利用者が極めて限定される状況”においては、ネットワークへ接続した状態での開発も可能とすべきである。	本ガイドラインは、ネットワークに接続して行うＡＩの開発を制限するものではなく、ネットワークに接続して行うＡＩの開発について、開発者に留意することが期待される事項を整理したものである。したがって、大学の研究室等において、ネットワークに接続してＡＩの研究開発を行う際には、本ガイドラインに留意することが期待される。 なお、ネットワークに接続してＡＩの研究開発を行う場合であっても、セキュリティが十分に確保されたサンドボックス等論理的に閉鎖された空間での開発は、本ガイドラインの対象とはならない。
「国際的な議論のためのＡＩ開発ガイドライン案」における開発原則の内容及び解説について		
５	ＡＩネットワーク化のセキュリティリスクを明確にすべきである。	ＡＩネットワーク化のセキュリティに関するリスクについては、報告書本体 53 頁に記載されているとおり、例えば、ＡＩシステムへのハッキング、偽装・なりすましによるＡＩシステムの犯罪への悪用等などが考えられる。本ガイドライン案の「セキュリティの原則」及びその解説も、このようなリスクを念頭に、作成されている。

6	「制御可能性の原則」という原則名称と内容が不一致であるので内容に即して「リスク評価の原則」とでも変更するべきである。	「制御可能性の原則」の解説においては、リスク評価に関し留意することが期待される事項を説明した上で、リスク評価を踏まえ、リスク管理に関し留意することが期待される事項（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））についても説明している。したがって、「制御可能性の原則」の内容は、リスク評価に限られるものではない。また、リスク評価のプロセスは、制御可能性のみならず、セキュリティや安全など他の原則においても期待されるので、「制御可能性の原則」に特有のものではない。
7	「プライバシーの原則」におけるプライバシーの定義如何。	プライバシーの概念については、国内外の判例や学説等においてもさまざまな定義が示されているため、本ガイドライン案においてプライバシーの定義を示すことを控えているが、「プライバシーの原則」の解説において、「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」と説明し、本原則において想定されるプライバシーの範囲を示している。
8	ガイドラインの基本理念である「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」が将来ＡＩによって損なわれる可能性があることが危惧されている。この危惧が生じないようにするためのガイドラインを加えるとよい。 ・　ＡＩは人間の道具に徹するべき。 ・　ＡＩに武器を持たせてはならない。 ・　ＡＩに自己増殖機能（勝手にコピーを作る）を持たせてはならない。	本ガイドラインは、「人間中心の智連社会を実現すること」を目的に掲げ、人間が主体となってＡＩを使いこなしていくという方向性を明確にしている。その上で、本ガイドラインの基本理念「人間がネットワーク化されたＡＩと共生することにより、その便益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」などを踏まえ、「倫理の原則」では、「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」と定めた上で、その解説において、自律型兵器に関する議論などを踏まえ、「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」と説明を加えている。 また、「制御可能性の原則」の解説において、脚注で掲げている「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」のほか、ＡＩの自己複製などによりＡＩが制御不能になるリスクなども見据え、制御可能性に関するリスク評価やリスク管理について、留意することが期待される事項を説明している。

９	「利用者支援の原則」の内容は、インターフェースに関する事項が中心となっているので、名称を「ユニバーサルデザイン原則」とでも変更するべき。	「利用者支援の原則」の解説では、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のみならず、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（いわゆる「ナッジ」）が利用可能であることに配慮するよう努めることも掲げている。したがって、「利用者支援の原則」の内容は、ユニバーサルデザインに関する事項に尽きるものではない。なお、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上では、インターフェースが重要な役割を果たすと考えられることから、「利用者支援の原則」の解説において、インターフェースに関する事項の説明が多くなることは適当であると考えられる。
１０	○　先端研究レベルでは、被験者の同意の上、プライバシー情報を用いた実験を行うことが多い。利用者の同意があれば、「プライバシーの原則」については、除外規定を受けることができるか確認したい。 ○　先端研究レベルでは、適切なインターフェースの設計などの段階で利用者支援が十分に行えないことが想定される。この場合、利用者の同意があれば、「利用者支援の原則」については、除外規定を受けることができるか確認したい。 ○　先端研究レベルでは、被験者に技術的特性について情報提供と説明を行うことで実験が成り立たない場合がある。利用者の同意があれば、「アカウンタビリティの原則」については、除外規定を受けることができるか確認したい。	開発原則は、開発者が遵守すべき基準を画一的に定めるものではなく、開発者が、自らの開発するＡＩシステムに用いられる技術の特性や用途に照らし、適切に留意して対応し、対応状況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指針を示すものである。したがって、例えば、実験段階において、被験者に適切な説明を行った上で、被験者の同意を得て、利用者支援及びアカウンタビリティ（技術的特性等に関する情報提供や説明）の程度や範囲を限定する場合には、当該被験者との関係においては、「利用者支援の原則」及び「アカウンタビリティの原則」に留意しているものと解するのが適当ではないか。また、一般に、本人の同意を得て個人情報を利用することがプライバシー侵害にあたるとは解されていないことから、開発者が、本人の同意を得て、個人情報を利用した実験を行う場合には、当該個人との関係においては、プライバシーの原則に留意しているものと理解するのが適当ではないか。

影響評価について		
１１	ＡＩが与える影響は単に対利用者といった単純なものではなく、社会システム全体に及ぶものであると考える。ＡＩを単に人間の機能の代替として見なすだけではなく、社会システムの変革をもたらすものと考えることが重要であり、ＡＩ開発・適用が社会システムに与える影響のアセスメントを並行して実施すること有意義なＡＩの実現に重要であると考える。	分野別評価においては、社会・経済システムの変革を念頭に、領域横断的なユースケースや領域が融合するユースケースを想定した評価を行っている。引き続き、社会・経済システムの変革を見据えて、ＡＩネットワーク化が社会・経済にもたらす評価を検討することとしたい。
１２	ＡＩネットワーク化に関する国際的な議論を進めるに際して、日本国内で開発され整備されるＡＩシステムを外国人が利活用するという状況にも十分に留意する必要がある。	日本国内で開発されたＡＩシステムを外国人が利活用することは十分に想定されることである。必要に応じて、分野別評価において、このような事項も考慮しつつ、インパクト及びリスクの評価を進め、その成果を開発ガイドライン及び利活用ガイドラインの検討に活用していきたい。
１３	有効なＡＩの活用方法として、物理的な方法を伴う物流、軽作業、重労働に用いるよりも、医療診断や裁判等への運用の方が相性がよいのではないか。下流の労働を解決するより、上流の労働を解決することが先決である。	これまでの検討から、様々な領域や分野におけるＡＩネットワーク化が社会・経済にもたらす影響に関する示唆を得ることができた。これを参考にしつつ、それぞれのインパクト（良い影響、便益）及びリスク等を勘案しながら、各領域、分野におけるＡＩシステムの導入が図られることが期待される。