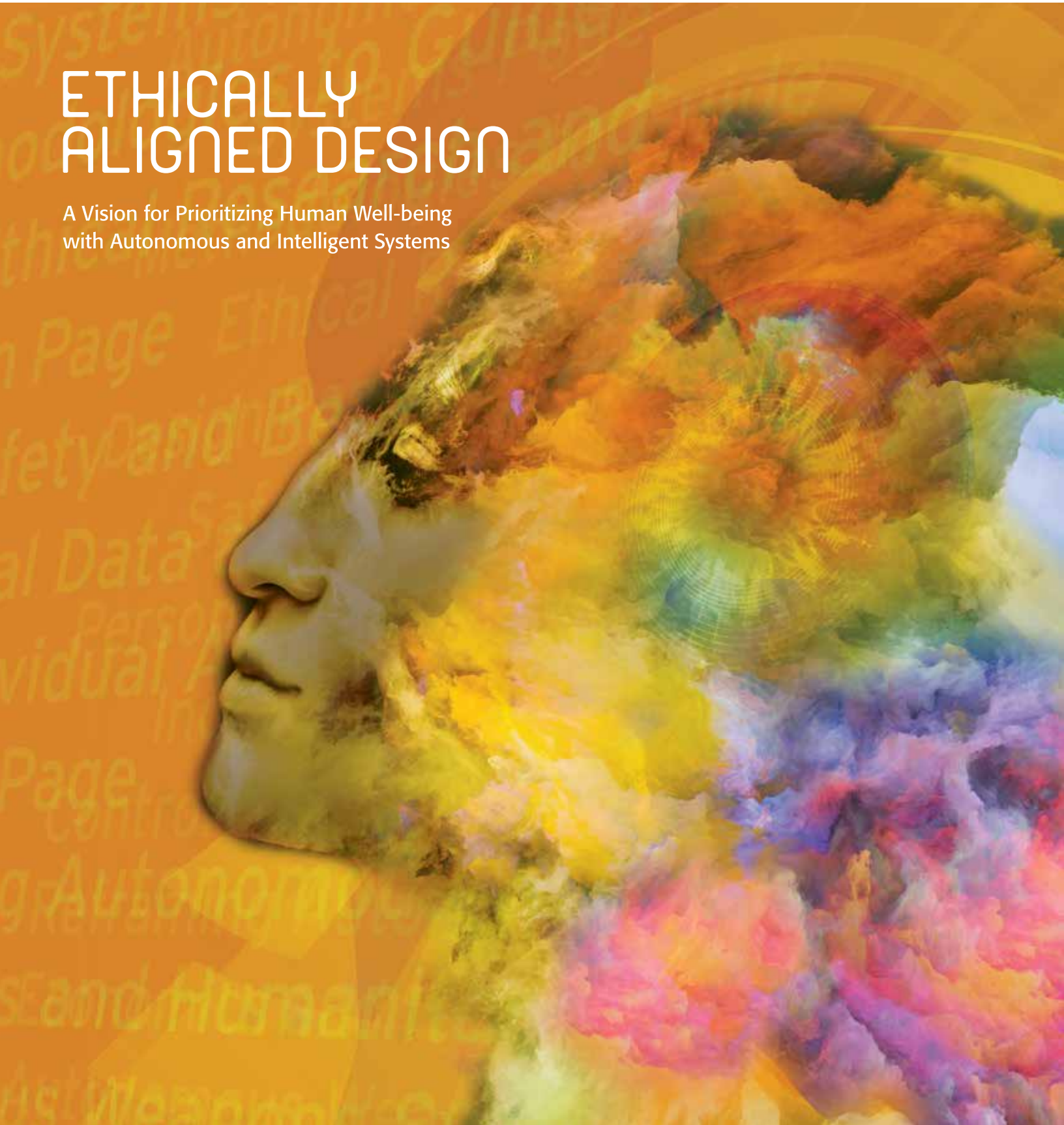
概要 - 第2版

IEEE
Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems



倫理的に配慮されたデザイン(Ethically Aligned Design)、第2版 – 意見募集

*『倫理的に配慮されたデザイン:自律インテリジェントシステム(A/IS)で人間の福祉を優先するためのビジョン』*第2版に対するパブリックコメントを募集します。この文書は、この種のシステムの開発にあたって科学技術者が倫理的な配慮を優先することを奨励するものです。

この文書は、自律インテリジェントシステムの倫理に関するIEEEグローバルイニシアティブ(以下、「IEEEグローバルイニシアティブ」)の委員会が作成しました。この委員会は、6つの大陸から集まった数百人の参加者で構成されています。関連する技術および人道主義の分野で、学術研究、産業、市民社会、政策・行政に携わる思想的リーダーたちが、タイムリーな問題についてコンセンサスを見出すために参加しています。

この文書の目的は次のとおりです。

- 文化的コンテキストの中で人間の福祉を優先するという、明確な価値観および倫理原則に適合する形で、自律インテリジェントシステムおよびテクノロジーを倫理的・社会的に実現するしくみを確立するにはどうすべきかについて、公的な議論を前進させる。
- 規格(IEEE P7000™シリーズ以降)ならびに関連する認証プログラムの立案にヒントを与える。
- これらの原則に即した国の政策および国際政策の出現を促す。

IEEEグローバルイニシアティブは、*『倫理的に配慮されたデザイン(Ethically Aligned Design)』*第2版に対するコメントを募集することにより、関連する科学技術コミュニティならびに一般社会のさまざまな声を集め、差し迫った倫理的・社会的な問題への幅広いコンセンサスを見出すとともに、これらのテクノロジーの開発と実装に関する推奨事項の案を募る機会とします。

*『倫理的に配慮されたデザイン(Ethically Aligned Design)』*に対するご意見は、電子メールで2018年3月12日までにお送りください。お寄せいただいたご意見は、*自律インテリジェントシステムの倫理に関するIEEEグローバルイニシアティブ*のウェブサイトで、2018年4月30日までに公開します。パブリックコメントの提出方法についての詳細は、提出に関するガイドラインをご覧ください。

この意見募集に対して寄せられたパブリックコメントは、2019年リリース予定の*『倫理的に配慮されたデザイン(Ethically Aligned Design)』*最終版に掲載される候補として、IEEEグローバルイニシアティブによる選考の対象となります。

詳しくはIEEEグローバルイニシアティブのウェブサイトでご確認ください。

ジャーナリストの方で、自律インテリジェントシステムの倫理に関するIEEEグローバルイニシアティブについて詳しい情報をご希望の場合は、IEEE-SA PRチームにご連絡ください。

倫理的に配慮されたデザイン(Ethically Aligned Design)、第2版 – 概要

*自律インテリジェントシステムの倫理に関するIEEEグローバルイニシアティブ*は、The Institute of Electrical and Electronics Engineers, Inc.(IEEE)のプログラムです。IEEEは、人間に恩恵をもたらすテクノロジーの進歩を目標とする、160カ国以上、420,000以上の会員を擁する世界最大の技術専門家団体です。IEEEグローバルイニシアティブは、250人以上の参加者で構成されています。6つの大陸に広がる自律インテリジェントシステムコミュニティで、学術研究、産業、市民社会、行政に携わる思想的リーダーたちが、この分野におけるタイムリーな問題についてコンセンサスを見出すために参加しています。IEEEグローバルイニシアティブの使命は、自律インテリジェントシステムの設計と開発に関わるすべてのステークホルダーが、倫理的な配慮を優先するための教育、訓練、奨励を通じて、人間に恩恵をもたらすテクノロジーの進歩を促すことです。

2016年、IEEEグローバルイニシアティブは、自律インテリジェントシステム(A/IS)、倫理、哲学、政策の各分野における上記コミュニティの意見を集約し、『*倫理的に配慮されたデザイン(Ethically Aligned Design)*』(EAD、第1版)を制作しました。 EADを出発点として、IEEEグローバルイニシアティブの勧告を反映し、現時点までに11個のIEEE P7000™規格ワーキンググループが設立されています。IEEEグローバルイニシアティブの目標は、『*倫理的に配慮されたデザイン(Ethically Aligned Design)*』とそれに基づくIEEE規格が、今後の技術開発のための重要な参照先となる知見および推奨事項を提供することです。以下、『*倫理的に配慮されたデザイン:自律インテリジェントシステムで人間の福祉を優先するためのビジョン』第2版*の概要を示します。

免責事項:『倫理的に配慮されたデザイン(Ethically Aligned Design)』*に記載された推奨事項は、IEEEの見解を表すものではなく、A/ISに関する専門的な指針となる知見を提供した委員会メンバーによる、正しい情報に基づく見解を表しています。IEEEまたはIEEE-SA Industry Connections Activityのメンバーは、本資料の誤りや脱落、あるいは直接的か否かを問わずいかなる点でも本資料に起因する損害について、かかる損害が事前に予見可能であった場合も含めて、一切の法的責任を負わないものとします。*

倫理的に配慮されたデザイン(Ethically Aligned Design)、第2版 – 概要

I. 目的

インテリジェントで自律的な技術システムは、日常生活の中で、人間による介入を減らすことを目的に設計されています。そのため、この新しい分野では、人間や社会に及ぼす影響をめぐって懸念が提起されています。現在の論点としては、望ましい影響に関して積極的に支持を表明する意見もありますが、プライバシー、差別、スキルの消失、経済的影響、重要インフラの安全保障、社会福祉に対する長期的な作用といったものに対する潜在的な有害性を根拠とした警告もあります。これらのテクノロジーの性質上、私たちの明確な価値観や倫理原則に適合して初めて、テクノロジーの恩恵が全面的に得られます。したがって、これらのテクノロジーに備わった技術以外の意味合いについて、正しい情報に基づいた対話や議論を方向付けるためのフレームワークを確立する必要があります。

II. 目標

これらのテクノロジーの倫理的な設計・開発・実装は、次の一般原則に従って行う必要があります。

- **人権:** 国際的に認知されている人権を侵害しないものであること
- **福祉:** 設計と使用において福祉のメトリックを優先すること
- **アカウンタビリティ:** 設計者および運用者が説明責任を確実に果たすこと
- **透明性:** 運用における透明性を確保すること
- **悪用に関する認識:** 悪用のリスクを最小化すること

II. 目標

個人データ権および個人のアクセスコントロール

基本的なニーズは、人々が各自の個人デジタルデータの使用をめぐって、アクセスを定義する権利と、正しい情報を得た上で同意を与える権利があることです。個人が独自のアイデンティティおよび個人データを監督するためのメカニズムと併せて、個人情報の一括化や再販によって生じる結果について、明確に知るためのポリシーおよびプラクティスが必要です。

経済効果による福祉の促進

誰もが手頃な費用で利用可能な通信ネットワークおよびインターネットを通じて、世界中のあらゆる場所で人々がインテリジェントで自律的な技術システムを利用し、恩恵に浴することが可能になります。これらのシステムは、より人間中心型の構造に向かって制度や制度的関係を大きく変え、人道主義的な問題や発展の問題に恩恵をもたらし、結果的に個人および社会の福祉を増大させる可能性があります。

アカウンタビリティに関する法的フレームワーク

インテリジェントシステムとロボット技術が収束し、部分的な自律性、固有の知的作業を遂行する能力、さらには人間のような外見という点で、人間をシミュレートした特性のあるシステムが開発されるようになっています。そのため、複雑・インテリジェント・自律的な技術システムの法的ステータスという問題は、どのようにアカウンタビリティを確保するか、この種のシステムが害を及ぼした場合

倫理的に配慮されたデザイン（Ethically Aligned Design）、第2版 – 概要

にどのように法的責任を配分するかという、より広範な法的問題と絡み合っています。考慮すべき一般的なフレームワークの例としては、次のものがあります。

- インテリジェントで自律的な技術システムは、適用可能な財産法制度の対象となるべきである
- 政府および業界のステークホルダーは、この種のシステムに委任してはならない意思決定および業務の種類を特定するとともに、そうした意思決定に対する人間の有効な支配力を確保するため、これらのシステムが害を及ぼした場合の法的責任の配分方法も含めて、規則および基準を採用すべきである

透明性と個人の権利

自己改良能力のあるアルゴリズムおよびデータアナリティクスによって、市民に影響を及ぼす意思決定を自動化することが可能になるとしても、透明性・参加性・正確性を義務付ける法的要件に、次のような目標を含める必要があります。

- 当事者、その弁護士、および裁判所は、政府その他の国家機関が採用するこの種のシステムで生成・使用されるすべてのデータと情報に対し、妥当なアクセス権を持たなければならない
- システムに組み込まれたロジックとルールを監督者が入手し、リスク評価や厳密なテストの対象とすることが可能でなければならない
- システムは意思決定の根拠となった事実と法律を記録した監査証跡を残すとともに、第三者の検証を受け入れる必要がある
- 一般社会は、投資を通じてこの種のシステムの倫理的判断を行っている主体、あるいはその判断を支持している主体を知っている必要がある

政策による教育と認知

効果的な政策によって、安全性、プライバシー、知的財産権、人権、サイバーセキュリティの保護・推進に対応するとともに、インテリジェントで自律的な技術システムが社会に及ぼす潜在的な影響について、一般社会の理解を促します。公共の利益に最大限に奉仕するには、次のような政策が必要です。

- 国際的に認知された法規範を支持・推進・有効化する
- 関連テクノロジーに関する労働人口の専門知識を発達させる
- 研究開発に対する指導力を発揮する
- 規制によって公衆安全と責任を確保する
- 関連テクノロジーが社会に及ぼす影響について社会を教育する

IV. 根拠

古典的な道徳規範

IEEEグローバルイニシアティブは、デジタル時代における人間の道徳という問題に対処するにあたり、2,000年以上に及ぶ古典的な道徳規範の伝統に基づいて、定着した倫理体系を詳しく調査しています。宗教から切り離された哲学の伝統も含めて、科学的アプローチと宗教的アプローチの両方を視野に収めています。自律性と存在論を定義する哲学的基盤の見直しを通じて、インテリジェントな技術システムに備わっているとされる自律的能力や、道徳を超越したシステムの道徳性について考察し、道徳を超越したシステムによる意思決定が、果たして道徳的結果をもたらし得るかどうかを問いかけています。

倫理的に配慮されたデザイン(Ethically Aligned Design)、第2版 – 概要

福祉のメトリック

拡張されたインテリジェンスと自動化によって人間に具体的な恩恵がもたらされるのであれば、その恩恵を表す明確な指標が必要です。成功の目安となる一般的なメトリックとしては、利益、労働安全、財務健全性などがあります。これらのメトリックは重要ですが、個人や社会の立場から見た福祉を、全範囲にわたって表わせるものではありません。心理的、社会的、環境的な要因も重要です。福祉のメトリックでは、このような要因も捉えられます。テクノロジーの進歩に起因する恩恵を、より総合的に評価することが可能になります。人間の福祉を損ないかねない、意図せぬマイナスの影響がないかどうかチェックする機会も得られます。逆の見方をすると、これらのメトリックは、インテリジェントな技術システムが人間の福祉を増大させる分野を特定するのに役立ち、社会的・技術的なイノベーションへの新たなルートとなる可能性があります。

自律システムへの価値観の組み込み

マシンが人間のコミュニティに準自律的な代理人として関与する場合、これらの代理人は、コミュニティの社会的・道徳的な規範に従うことが期待されます。この種のシステムに規範を組み込むには、導入先となるコミュニティの明確な描写が必要です。さらに、同じコミュニティの中でも、技術の具体化表現の種類に応じて、それぞれ異なった規範の集合が求められます。システムの導入先となる特定のコミュニティの規範、とりわけシステムの設計目標であるタスクに関連した規範を明らかにすることが、最初のステップです。

倫理研究と設計の指針となる方法論

人間の福祉と自由を強化・拡大するインテリジェントな技術システムを開発するには、価値ベースの設計方法論により、マシンが人間に奉仕すべきであってその逆ではないという認識に従って、人間の進歩を技術システム開発の中核に位置付けます。社会的コストと、組織の経済的価値を増大させる利点の両面から評価することのできる、持続可能なシステムを開発するため、システム開発者がこのような価値ベースの設計方法論を採用する必要があります。

V. 将来的なテクノロジーに関する懸念事項

自律兵器の見直し

身体的危害を引き起こすよう設計された自律システムは、従来型の兵器や、危害を及ぼすことを目的としない自律システムと比べて、さらに別の倫理的次元があります。この倫理的次元には、少なくとも次のものが含まれます。

- 兵器システムに対する人間の有意な統制力を確保する
- 自動兵器の設計における監査証跡を残し、アカウンタビリティと統制力の保証に役立てる
- システムが行った推論と意思決定を、人間のオペレーターに対して高い透明性で分かりやすく説明できる、適応型の学習システムを含める
- 自律システムの運用を担当する、明確に識別可能な人間のオペレーターの訓練を行う

倫理的に配慮されたデザイン（Ethically Aligned Design）、第2版 – 概要

- 自律機能の振る舞いを、オペレーターにとって予測可能なものにする
- これらのテクノロジーの開発者に、自分が行う作業の暗黙的な意味合いを確実に理解させる
- 危害を及ぼすことを目的とした自律システムの開発に適切に対応する専門的な倫理規定を策定する

汎用人工知能（AGI）および人工超知能の安全性と恩恵

インテリジェントで潜在的に自己改良能力のある技術システムの開発と使用は、他の強力なテクノロジーと同じく、悪用または拙劣な設計のどちらに起因するにせよ、相当なリスクを内包しています。一部の理論によれば、システムがAGIに近付き、それを超えるようになると、システムによる予想外の振る舞いが危険性を増し、修正がますます困難になります。AGIレベルのアーキテクチャは必ずしも、人間の利益に合致させることが可能なわけではないと予測されます。そのため、アーキテクチャが能力を増すにつれ、各アーキテクチャがどのように動作しているかを慎重に判別する必要があります。

感情コンピューティング

感情は、知能の中核をなす側面の1つです。私たちの生活全体を通じて、怒り、恐れ、喜びといった衝動や感情が行動の根拠になる場合が少なくありません。インテリジェントな技術システムを通じて、あらゆる状況で可能な限り人間を支援するためには、人間の社会に参加する、あるいは社会を支援するために存在する人工物が、人間の感情的体験を増幅または減衰させるという実害を及ぼすことがあってはなりません。現在すでに一部のシステムに導入されている、感情合成の初歩的なバージョンでさえも、これらのシステムに対する政策立案者や一般社会の受け止め方に影響しています。

混合現実

混合現実が私たちの仕事、教育、社会生活、商取引の中でさらに一般化すれば、この種のテクノロジーによって、アイデンティティや現実に対する私たちの概念が変わっていく可能性があります。このような混合現実の世界に備わったリアルタイムパーソナライゼーションの能力は、個人の権利や多面的なアイデンティティに対する統制力をめぐって、倫理上の問題を提起します。ヘッドセットを装着する必要もない、より微妙で統合的な知覚強化に向かってテクノロジーが進化していくのであれば、なおさらです。