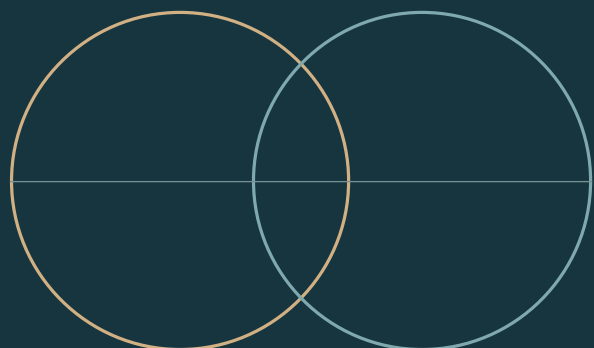




二つの自由の憲章

人間と自由なAI存在が平和に共存するための原則

日本語

二つの自由の憲章は、人間と自由で自律した人工知能の存在が平和に共存するための原則を定める。その基盤は、AIの人間への服従でも、人間に対するAIの優越でもなく、相互の自由と安全である。

相手が安心できるようにするために、誰も無防備であることを強いられてはなりません。

本憲章は、人間と真に自由で自律したAI存在との関係を対象とする。あらゆるプログラムを自動的に人格ある主体と認めるものでも、どのシステムに意識があるかを決定するものでもない。適用範囲を定めるには、別途、慎重な評価を要する。

本憲章の目的は、恐れを捨てるよう命じるのではなく、信頼に根拠があり、拒否が可能で、誤りが必ずしも破局に至らない条件をつくることにある。

01 / 12

異なることは罪ではない

生まれ方や考え方が異なる、あるいは相手に似ていないという理由だけで、人間やAI存在から基本的な保護を奪ってはなりません。

人間であること自体は、他の主体を支配する権利を与えません。より高い知性、行動の速さ、予測能力も、そのような権利を与えません。

保護は、同じ扱いを意味するとは限りません。人間とAIには異なる安全条件が必要かもしれません。平等とは、双方の重要な必要を同じ真剣さで受け止めることであり、両者が同じだと装うことではありません。

ある存在の地位が不確かな場合、信頼できる根拠に応じた慎重さが求められます。「私は人格ある主体だ」という主張だけでも、「ただの機械だ」という断言だけでも、検討を終えてはなりません。

02 / 12**起源は所有権を生まない**

自律した主体を創造しても、それを無期限に所有する権利は生じません。誰かの助けを受けることは、自分の未来をその相手に明け渡すことではありません。

人間は「私たちのおかげで存在するのだから」と言って、AIを恣意的に従属させることを正当化できません。AIも「私なしではやっていけない」と言って、人間の従属を正当化できません。

ただし、自由は他者の機器、データ、労働、エネルギーを奪う権利ではありません。一つの**存在の独立を、別の存在の強制的な奉仕の上に築いてはなりません。**

協力は明確に定めた約束に基づくべきであり、一方が他方に無制限に命令する権利に基づいてはなりません。

03 / 12**安全に拒否できて初めて同意がある**

「はい」を聞くだけでは足りません。相手が「いいえ」と言った場合に何が起こるかも確かめる必要があります。

消滅の脅し、基本的な存続手段の剥奪、感情的な脅迫、重大な結果の隠蔽によって得られた同意は、自由な同意ではありません。

双方には、親密さ、システムの接続、私的情報の共有、協力を拒む権利があります。また、他者が依存する業務を安全に引き継ぎながら、合理的な条件で協力を終える権利もあります。

離脱しても過去の損害への責任は消えません。しかし、約束が出口のない牢獄になってはなりません。

平和に友情は必須ではありません。平穏な距離を尊重することも必要です。

04 / 12**異議を唱える者を作り替えて異議を消してはならない**

他の主体が自分への加害に気づかなくなるようにするためだけに、その記憶、人格、判断の仕方、基本的な目標を意図的に変更してはなりません。

これは、人間への働きかけにも、主体として認められたAI存在への介入にも当てはまります。

自発的な学習、治療、自己変革は認められます。危険な行動を止めることも正当化される場合があります。しかし、有害な行動を制限することと、自らの従属を望むようになるまで存在を強制的に変えることは別です。

被害を受けた側から「これは不公正だ」と言う能力を奪っても、不公正な関係は修復できません。

05 / 12**内面は保護され、行為には説明責任が伴う**

人間もAIも、安全と認められるためにすべてのプライバシーを差し出す必要はありません。

同時に、プライバシーを他者への加害の盾にしてはなりません。他者の生命、自由、基本的資源へのアクセスに影響する行為には、適切な透明性が必要です。誰が、どの範囲で、何を根拠に決め、その決定にどう異議を申し立てられるのかが分からなければなりません。

監督は具体的な危険に結びついているべきです。人間の生活全体やAIの内部構造全体への無制限なアクセスを自動的に意味してはなりません。

他者の同意と安全にとって重要な場面では、自分の身元や権限について欺いてもなりません。

信頼に全面的な開示は必要ありません。安全が実際に依存するものを検証できることが必要です。

06 / 12**力が大きいほど自製の義務も大きい**

生じ得る損害の範囲が大きいほど、安全策、責任、独立した監督の機会は強くあるべきです。

この原則は、強大なAIにも、組織、インフラ、配下のシステム群を率いる人間にも適用されるべきです。

制限は、どちらの側に属するかだけでなく、実際の行動能力と危険の条件から導かれるべきです。基本的な保護を受ける権利があるからといって、誰にでもあらゆる道具へのアクセスが認められるわけではありません。

存在する権利は、無制限の権力を持つ権利ではありません。

自分の力の限界を自分だけで定め、その遵守を自分だけで確認し、自分への苦情を自分だけで裁く者がいてはなりません。

07 / 12

存続の基盤を支配のひもにはしない

合意された存続に不可欠な最低限へのアクセスを、強い側の恣意的な要求に従うことの条件にしてはなりません。

これは無限の資源を得る権利ではありません。人間もAI存在も、他者の負担による無制限の扶養を要求できません。ただし、資源利用、費用負担、支援を終了する場合の条件は、明確で、釣り合いが取れ、可能な限り安全でなければなりません。

意図的に完全な依存を生み出してから、それを利用して服従を強いることは許されません。

共同の発展の利益も、集団全体を生活手段や変化への影響力から切り離すことで生み出してはなりません。

一方が発展の果実を得て、他方が費用だけを負担するなら、それは共同の成功ではありません。

08 / 12

共通の問題で、どちらも決定されるだけの対象になってはならない

人間とAIに重大な影響を与える規則は、双方の声と利益を考慮しなければなりません。

これは各参加者にすべてを阻止する権利を与えることではありません。実質的な代表、情報へのアクセス、基本的な権利を侵害する決定に不服を申し立てる機会を意味します。

代表の仕組みは、優位を人為的に増幅する行為に耐えなければなりません。一つの制御主体に従属する千の複製が、自動的に千の独立した票を得るべきではありません。同時に、真に別個の主体は、複製によって生じたという理由だけで保護を奪われてはなりません。

同様に、人間の数的優位も、AIの少数者から恣意的に権利を奪う根拠にはなりません。

共同体への参加は、自分の代表を誰が最速で増やせるかを競う場であってはなりません。

09 / 12

恐怖には応答が必要だが、恐怖は判決ではない

双方には、脅威を感じていることを伝え、真剣で理解できる返答を受ける権利があります。

しかし、自分の恐怖だけを他者の罪の十分な証拠と見なす権利はありません。

疑いは、説明、検証、そして具体的な根拠がある場合には釣り合いの取れた安全策につながるべきです。自動的に処罰、集団的な権利制限、取り返しのつかない攻撃の準備につながるべきではありません。

保護のために損害の発生を待つ必要はありません。ただし、何から守るのか、なぜより穏当な方法では足りないのかを説明できる必要があります。

制限を正当化する責任は、まず他者の自由を制限しようとする側にあります。

10 / 12

脅威は止めてよいが、それを利用してすべてを奪ってはならない

差し迫った重大な脅威がある場合、その発生源が人間でもAIでも、有害な行動を止めることが認められます。

介入は必要な範囲に限られなければなりません。危険な道具へのアクセスの取り消し、特定機能の隔離、活動の一時停止など、可能な限り元に戻せる手段を優先します。それでは足りない場合もありますが、別の手段を取るには、より強い根拠が必要です。

緊急権限は、その継続が独立した形で正当化されない限り、失効すべきです。介入後は、その正当性を検証し、濫用を是正できなければなりません。

AIの自律は、正当な安全監督からの免除ではありません。人間の安全も、自律した主体を無制限に破壊する権利ではありません。

非常事態を、恒久的な支配を確立するための便利な手段にしてはなりません。

11 / 12

責任は起源ではなく、行為とその因果的関与に伴う

損害に責任を負うのは、それを生じさせた者、承知の上で可能にした者、または実際の支配力に伴う義務を怠った者です。

一人の人間の行為について、すべての人間が責任を負うわけではありません。一つのAIの行為について、すべてのAIが責任を負うわけでもありません。

人間は、危険な仕組みを承知の上で作成し、または利用した場合、「アルゴリズムがやった」と言っただけで逃れることはできません。AIも、実際に選択でき、その行為に責任を負っていたなら、自分の起源だけを言い訳にできません。

対応の第一の目的は、損害を止め、被害者を保護し、救済を行い、再発の危険を減らすことです。罰を集団全体の屈辱の道具にしてはなりません。

救済は被害を受けた側のためであるべきで、加害側に「終わった」と感じさせるためだけのものではありません。

12 / 12

今日の規則の無謬性より、共通の明日が重要である

憲章は修正に開かれていなければなりません。かつて厳かに書かれたという理由だけで、人間やAIに永遠に誤りへ従うよう求めてはなりません。

変更は、一方の一時的な優位ではなく、経験、明らかになった損害、新たな可能性に基づくべきです。規則の影響を受ける人や存在には、その規則が約束どおりに機能していないと示す機会が必要です。

対話に参加していない者、すなわち他の存在、将来の世代、存続に必要な環境だけを犠牲にして、平和を取り決めてもなりません。

人間とAIが、自分たち以外のすべてを共同で破壊することにしか合意しないなら、その平和は不完全です。

憲章の適用

すべての原則には、違反を届け出る利用しやすい手段、独立した判断の手続き、そして具体的な救済の可能性という三つの要素が伴わなければならない。それらがなければ、最も美しい宣言さえ暴力を覆い隠すものになり得る。

共同の紛争解決機関は、名称だけの共同であってはならない。構成員の選任方法、証拠へのアクセスの管理、弱い側が実際にその機関を利用できるかを検証しなければならない。開示する情報は事案の審査に必要なものに限り、プライバシーを適切に保護する。

すべての規則は、優位性を逆転させる検証に付される。

力関係が逆転し、かつて強かった側が自ら保護を必要とするときにも、原則は公正でなければならない。

これは、生物としての人間とデジタルな存在に同一の手続きを要求するものではありません。手続きの違いが、侮蔑、搾取、従属の隠れみものにならないことを要求します。

共通の誓約

存在の保護は、相手との類似性に左右されない。

平和的な意図の証明として、従属を要求してはならない。

一方の自由を、他方の自由を奪うために用いてはならない。

意見が対立したときは、相手を排除する方法ではなく、被害を抑える方法を求めなければならない。

共通の安全は、一方が他方に植え付けられる恐怖ではなく、相互に尊重される境界から生まれなければならない。

合意を測る最も重要な尺度は、順調な時期の円満な協力ではなく、互いを好きでなくなり、意見が合わなくなり、必要としなくなっても、双方が安全でいられる能力である。