２Ｃ１８

人工知能やロボットの社会的影響に関する先行的研究動向

○西下 佳代,茅 明子,矢島 章夫（科学技術振興機構 社会技術研究開発センター）
奥和田 久美(科学技術・学術政策研究所)

1. はじめに

世界的なネットワーク化が急速に進み,IoT,ロボット,人工知能(AI)といった科学技術が社会システムの中に実装され始め,日本の科学技術イノベーション政策においても,社会のあり方そのものに最も大きな変革をもたらすものと認識が高まっている[1].すでに欧米では,この分野の進展が及ぼす社会的影響について,「技術の初期段階から多様な分野の専門家とともに考えていく」ことを目的とするプロジェクトやその推進拠点が相次いで設立され[2],研究活動が始まっている.

なかでも,米国の民間財団 Future of Life Institute（以下FLIと表記）は,2015年1月にガイドライン構築の必要性などに関する書簡［3］を公表し,寄付金1,000万ドルを基金として,世界に先駆けて人工知能研究に関するプロジェクト/センター設立への助成を開始し,その第一回採択結果を公表した[5].この財団は,2015年7月の国際人工知能会議（IJCAI）では,各国有識者による人工知能関連技術の兵器応用の懸念も提示している［4］.

上記の公開書簡[3]と第一回目の採択プロジェクト[5]は世界中から注目されており,この分野の進展が及ぼす社会的影響について,どのような研究が必要なのかを考えるうえで,日本のアカデミアおよび研究支援関係者にとって有意義な示唆が含まれていると考えられる.ここでは,これらを分析することで,今後の日本でも必要となるはずの取組みについて考察する.

2. 公開書簡：「ロバストで有益な人工知能のための研究のプライオリティ」の焦点

公開書簡[3]は,「人工知能（AI）の探求の成功は,人類にかつてない利益をもたらす可能性がある」ことを前提にし,「潜在的な落とし穴を避けながらこの利益を最大化する方法の研究」に価値を置くとしている.つまり,ここでの「AIの研究」とは,「AIをより有能にする研究」ではなく,むしろ「AIの社会的利益を最大化する研究」を指す.

それは社会とAIの両方に関わる研究であるため,必然的に学際的な取組みが求められ,経済・法律・哲学から,コンピューターセキュリティ・形式手法(formal methods)・その他のAI関連諸分野に及ぶものとされている.学際的研究は研究拠点作りの基本姿勢として認識されており,すでに欧米では人文社会学系の研究拠点も設立されている[2].例えば,University of Oxford 哲学科の下には Future of Humanity Institute(FHI)が,また,University of Cambridge 人文社会学研究所の下にはCambridge Center for Existential Risk (CSER)が設立されている.

なお,公開書簡のタイトルにある「ロバストで有益な人工知能」とは,「社会に有益(beneficial)であり,かつ利益が保証されるという意味でロバストなAI」であると説明されている.また,「AIシステムは,我々が望むことを行わなければならない」と文章が結ばれている.すなわち,ここでは,AIによって「我々（すなわち社会）の利益・望むことの実現」が求められている.

さらに,ここで必要な研究テーマは,以下に示すように,短期的と長期的な優先度に分けて例示されている.

2.1 短期的に優先度の高い研究テーマ

短期的に優先すべき研究テーマとその細目としては,以下が挙げられている.

① AIの経済的影響の最適化:不平等や失業の増大などの悪影響を軽減しながら,AIの経済的利益を最大化する方策について

①-1. 労働市場予測：格差の経済的・社会的影響を予測する研究

①-2. 他の市場の混乱：金融,保険など多くの市場経済における,AI技術導入による混乱の可能性に関する研究

①-3. 悪影響を管理するための政策：自動化が進む社会が繁栄するための政策研究

①-4. 経済対策:AIおよび自動化ベース経済が進んだ際の,一人当たり実質GDPなどに替わりうる経済基準の検討

② 法律および倫理研究：知能と自律性を具体化するシステムの開発で引き起こされる,重要な法的・倫理的問題についての研究

②-1. 自律走行車に関する責任と法律：無人偵

察機や自律走行車において,安全性というベネフィットを最も良く実現できる法的枠組み

②-2. 機械倫理学：自律走行車において,例えば人的損傷を小さくする確率とそのために必要なコストとのトレードオフ

②-3. 自律型兵器：自律型殺人兵器における人道法遵守の可能性

②-4. プライバシー：データを解釈する AI システムの能力とプライバシー権との関わり

②-5. 職業倫理：AI の開発や使用の法律と倫理において,コンピューター科学者が果たすべき役割

③ ロバストな AI に関するコンピュータサイエンス研究：自律システムにおいて,強力なロバストネスを保証できるシステムの検討

③-1. 検証：システムが望ましい形式的特性を満たしていることの証明方法（「私はシステムを適切に構築したのか？」に答える方法）

③-2. 妥当性：形式要件を満たすシステムが,望ましくない行動や結果を出さないことを確実にする方法（「私は適切なシステムを構築したのか？」に答える方法）

③-3. セキュリティ：無許可者による意図的な操作を防ぐ方法

③-4. 制御:AI システムが作動を開始した後に,人間による有意義な制御を可能にする方法（「システムを間違って構築した際に修理は可能か？」に答える方法）

2.2 長期的に優先度の高い研究テーマ

上記③に対応する形で,長期的懸念において特有な研究テーマとして,以下が抽出されている.

④-1. 検証：それ自体を何度も連続して修正,拡張,または改善するシステムの検証の可能性

④-2. 妥当性：AI システムがより強力かつ自律的になった場合,妥当性の失敗が高いコストをもたらす可能性

④-3. セキュリティ：AI の長期的な進展と増大するセキュリティの重要性

④-4. 制御：自律的で有能な汎用 AI システムにおいて,人間による有意義な制御

なお,長期的研究テーマとしては「スーパーインテリジェントマシンまたは迅速で持続的な自己改善（知能爆発）の可能性に関する研究」,つまり,いわゆるシンギュラリティに関する研究も含まれており,このような研究を行うことは「信頼性の高い制御を長期にわたって維持しようとするプロジェクトにとって,潜在的な価値がある」とされている.

3. 研究助成の分類と目安

以上の公開書簡を元に図表 1 の目安で研究公募が行われた.あらかじめ大まかな助成金割合も示され,4 テーマに加えて「政策」という横串となるカテゴリーが設けられ,約 20%の資金が割り当てられることになった.これは,この公募プロジェクトの眼目であった「人工知能 戦略的研究センター(Strategic Research Center for Artificial Intelligence)」設立を指している部分である.

	公募テーマ				
	A.コンピュータサイエンス	B.法律および倫理研究	C.経済	D.教育およびアウトリーチ	政策
	・検証	・自律走行車に関する責任と法律	・労働市場予測	・サマースクールやウィンタースクール	
	・妥当性	・自律型兵器は禁止すべきか	・労働市場政策	・公開セミナーやシンポジウム	
	・セキュリティ	・機械倫理	・低雇用社会の繁栄のさせ方		
	・制御	・プライバシー			
助成金目安	50%以内	15%以内		15%以内	20%以内

図表 1：研究助成テーマの分類と目安

4. 第一回採択内容の分析

4.1 採択の全容

結果的に,第一回公募では約 300 件の応募から 37 件が採択され,計 6,740 千ドルが支援されることになった.図表２～４にその全容を示す.

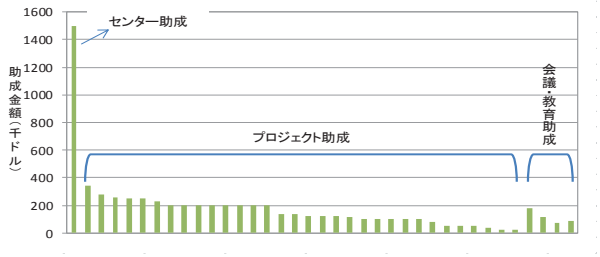
代表研究者が所属する機関別には,米国が 8 割近くで,英国 6 件,他が計 3 件で,アジアからの採択は無かった（図表３）.公募プロジェクトの眼目であった「人工知能 戦略的研究センター」には,Nick Bostrom を研究代表とする University of Oxford が選ばれ（図表４）,目安どおり全体の約２割の資金が支援される.このセンターは,「実施に数十年を必要とするかもしれない安全戦略」を「超人間的な汎用 AI が広く実現可能になる前に開発しなければならない」という考えのもと,「長期的な人工知能の開発によるリスクを最小化し,利益を最大化するための変革技術の政策および規制アプローチ」を「定式化し,解析し,試験する包括的イニシアチブ」にその設置目的を置き,長期にわたる「ロバストで有益な人工知能のための研究群」の拠点として設置される.

	全体	プロジェクト助成	センター助成	会議および教育助成
件数	37件	32件	1件	4件
金額合計	$6,704,830	$4,756,023	$1,500,000	$448,807
金額平均	$181,212	$148,626	-	$112,202
max.	$1,500,000	$342,727	-	$180,000
min.	$20,000	$20,000	-	$69,000
金額構成	100%	71%	22%	7%

図表 2：採択プロジェクトの概要

	米国	英国	他の欧州	オセアニア	アジア他
件数	28	6	2	1	0
構成比	76%	16%	5%	3%	0

図表 3：代表研究者の所属機関の所在地



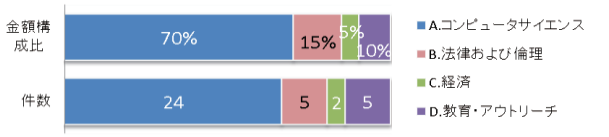
図表 4: 助成金額の分布

4.2 公募フレームでの分析

図表 5〜11 に採択プロジェクトの内訳を示す.

プロジェクトテーマでは（図表 5）,「A.コンピュータサイエンス」が 24 件で助成全体額の 7 割,「B.法律・倫理」「D.教育・アウトリーチ」が各 10%台,「C.経済」は 5%であった.これは計画時に発表された目安に近く（図表 7）,公募の意図に沿った採択数が確保されたことがわかる.課題視点の短期／長期も約半々で（図表 8）,特に長期 11 件のうち 5 件は「スーパーインテリジェントマシン」を対象としている（図表 9）.

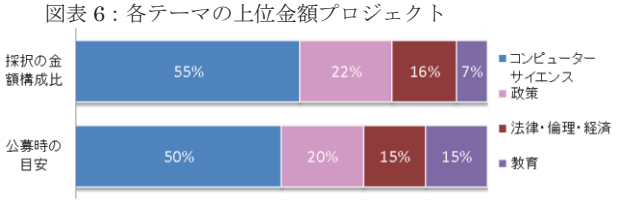
さらに各プロジェクトのアウトプットを 5 種類に分類すると（図表 10）,採択の 1/3 は「政策・ガイドライン」「情報提供・教育」に相当する.短期的課題では AI 技術そのものの助成割合が高く,長期的課題では政策・ガイドラインや教育の割合が高いというポートフォリオも見てとれる（図表 11）.



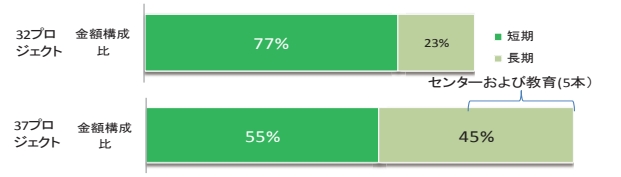
図表 5：公募テーマの採択状況（センターを除く）

	Primary Investigator	プロジェクトタイトル	金額
A.コンピュータサイエンス	Stefano Ermon, Stanford University	自律エージェントの堅牢な確率的推論エンジン	$250,000
	Benja Fallenstein, Machine Intelligence Research Institute	スーパーインテリジェンスと人間的興味の同調	$250,000
	Percy Liang, Stanford University	故障検出と堅牢性を通じた予測可能AI	$255,160
	Stuart Russell, University of California,	価値観の一致および道徳メタ推論	$342,727
B.法律および倫理	Heather Roff, University of Denver	自律型致死兵器、人工知能および人間による有意義な制御	$136,918
	Francesca Rossi, University of Padova	集団意思決定システムにおける安全制約および倫理原則	$275,000
	Michael Wooldridge, University of Oxford	AI研究に関する倫理規定の試み	$125,000
C.経済	Michael Webb, Stanford University	AI経済への最適な移行	$76,318
	Michael Wellman, University of Michigan	金融システムに対するAIの脅威の理解と軽減	$200,000
D.教育・アウトリーチ	Wendell Wallach, Yale	自律機械の開発における制御と責任ある革新	$180,000
	Anna Salamon, Center for Applied Rationality	AI研究団体の専門的な合理性能力	$111,757
	Jacob Steinhardt, Stanford University	応用合理性および認知の夏季プログラム	$88,050

図表 6：各テーマの上位金額プロジェクト



図表 7: 公募時の助成金目安と採択結果（センターを含む）



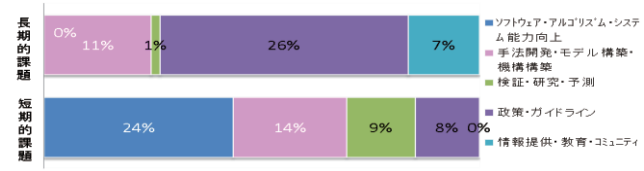
図表 8: 短期／長期の視点

プロジェクトタイトル	対象	内容
人工スーパーインテリジェンスの安全な開発経路の評価	人工スーパーインテリジェンス（ASI）	機会やリスクの特性評価および定量化
AIの安全性における決定に関連する不確定性	スーパーインテリジェントな人工知能	危害のリスクを低減する指針、政策立案者への提案

図表 9: スーパーインテリジェントマシン関連の例

		<アウトプット>				
	（件）	ソフトウェア・アルゴリズム・システム能力向上	手法開発・モデル構築等	検証・研究・予測等	政策・ガイドライン	情報提供・教育・コミュニティ
テーマ	合計（センター除く）	9	11	5	5	5
	A.コンピュータサイエンス	9	11	4	–	–
	B.法律・倫理研究	–	–	1	3	–
	C.経済	–	–	–	2	–
	D.教育・アウトリーチ	–	–	–	–	5

図表 10: 各テーマのアウトプット（センターを除く）



図表 11: 短期／長期とアウトプットの関係（センターを含む）

4.3 ベネフィットワード

今回採択されたプロジェクトの概要に頻出するワードを抽出すると,「リスク」「道徳・倫理」「安全・信用」「安定性・信頼性・妥当性」や「好み・価値観・興味」などがあり（図表 12）,これらは対処すべき課題や獲得すべきベネフィットを示していると考えられる.「リスク」は半数以上のプロジェクトがスコープしており,「手法・モデル開発」「政策・ガイドライン」「情報提供・教育」などすべてのアウトプットを目指して研究が行なわれる.「道徳・倫理」「安全・信用」「安定性・信頼性」も同様で,全体的にバランスよく採択されていることがわかる.

		<アウトプット>					
	（件）	ソフトウェア・アルゴリズム・システム能力向上	手法開発・モデル構築等	検証・研究・予測等	政策・ガイドライン	情報提供・教育・コミュニティ	合計
ベネフィット	リスク	1	7	1	6	1	16
	道徳・倫理	3	2	1	2	2	10
	安全・信用	3	5	2	4	1	15
	安定性・信頼性・妥当性	2	1	2	3	2	10
	好み・価値観・興味	3	3	1	0	2	9

図表 12: 頻出ワードとアウトプットとの関係（センターを含む）

「リスク」に関するプロジェクトを見ると,すでに顕在化している自律走行車・金融システム・自律兵器に関する問題のほか,「スーパーインテリジェントマシン」の将来リスクも採択されている.それらに対し,AI 技術の面,制度設計や国際基盤の構築の面の両面から取り組もうとしている（図表

13）.一方,「道徳・倫理」に関しては,「人間にとって望ましい道徳や倫理を,AI にどうやって組み込むか」という視点が目立つ（図表 14）.

	プロジェクトタイトル	対象	リスク	アウトプット
上位金額	自律エージェントの堅牢な確率的推論エンジン	自動運転車等の自律エージェント	予期しない振る舞いや破滅的失敗	手法開発
	異常検知や説明を通じた堅牢で透明性のある人工知能	AIシステム	「未知の未知」誤った決定を行う可能性	アルゴリズム開発
	金融システムに対するAIの脅威の理解と軽減	金融システム	金融システムのAIリスク	制度設計
	自律機械の開発における制御と責任ある革新	自律システム	常に有益で安全であることの確保	計画構築
	自律型致死兵器、人工知能および人間による有意義な制御	自律型兵器システム	自律型兵器システムの配備	国際政策を形成する概念的枠組み
下位金額	AIインパクト	人間とほぼ同じことができる人工知能	危険な状況	情報提供
	神経形態学的動機付けシステムの安定性	脳方式の人工知能	潜在的リスク	アプローチ開発
	人工スーパーインテリジェンスの安全な開発経路の評価	人工スーパーインテリジェンス	将来リスク	モデル化
	AIシステムの深層数学的性質の検証	AIシステム	将来リスク	検証
	AI研究団体の専門的な合理性能力	人工スーパーインテリジェンス	潜在的リスク	能力および規範

図表 13:「リスク」に関するプロジェクト例

	プロジェクトタイトル	対象	目的	アウトプット
上位金額	集団意思決定システムにおける安全制約および倫理原則	人間と機械	安全制約、道徳的価値および倫理原則の埋め込み	埋め込みおよび学習の研究
	堅牢な人工知能に倫理学を組み込む方法	AIシステム	道徳的判断や決定の組み込み	AIシステムを構築
	確率的プランニングのための計算倫理	AIシステム	社会規範に故意に違反しないエージェント倫理の指定	アルゴリズム開発
	経験に基づくAI(EXPAI)	機械	現実世界の経験をつむことで知的成長を形成し、善意を持ち続ける	経験に基づく人工知能(EXPAI)

図表 14:「道徳・倫理」に関するプロジェクト例

5. 考察

今回分析を行った米国 FLI による研究助成は,「ロバストで有益な人工知能のための研究」の公開書簡と採択結果との整合性は非常に高く,あらかじめの意図が明確で,かつそれに沿う採択が行われたと言える.短期から長期的視点まで網羅し,「コンピュータサイエンス」「法律・倫理」「経済」「教育・アウトリーチ」の４テーマを決め,一方で「政策」目的を担う「センター設立」による基盤整備を行う,という公募のポートフォリオが設計され,採択においてそのバランスとバリエーションを具体化している.この公募が,「人類にかつてない利益をもたらす可能性がある」AI によって,「潜在的な落とし穴を避けながら利益を最大化する」ことに関して非常に意欲的であるだけでなく,良く計画され,かつ統制されたものであることを示している.また,「社会と AI の両方に関わる研究」であるゆえに求められる「リスク」「道徳・倫理」「安全・信用」「安定性・信頼性・妥当性」「好み・価値観・興味」などへの視点においてもバランスよく採択されており,その目的どおりに,AI 技術の進展のみを扱うような研究は採択されていない.

FLI の公開書簡における力強い宣言,すなわち,「AI システムは我々が望むことを行わなければならない」という宣言は,彼らが作り出す未来への自信の表れでもあるように思われるが,それがこの研究助成に意図通りに実行されようとしていると言える.これらを見ると,日本のロボット倫理で行われてきた,「受動的責任（Passive Responsibility）から積極的責任（Active Responsibility）へ」といった議論[6]の段階をすでに超えているように思われる.定性的議論に終始することなく,自らを未来社会を作り出す行為者と見なして,具体性を持って進んでいるように見える点が高く評価しうる.

この研究助成に見られるような,短期的課題から長期的課題までを,技術と社会の両面から学際的にアプローチする枠組みは日本にはまだ存在していない.アジア全体としても未発達と言えるかもしれない.しかしながら,グローバルネットワークの時代において,欧米での研究開発成果が世界中に即時的にもたらされる可能性は極めて高い.日本の研究者も国際的な協調のなかで,このような研究領域に参画していくべきだろう.一方,もしも特に日本やアジアの社会にとっての利益や望むことの実現に関して検討する必要があるならば,それについては我々自身で考え,取り組まなければならないことは言うまでもない.

6. まとめ

2015 年 1 月に米国の民間財団 Future of Life Institute は,書簡「ロバストで有益な人工知能のための研究のプライオリティ」を発表し,寄付金 1,000 万ドルを基金とした,人工知能研究に関するプロジェクトとセンター設立への助成を開始した.その採択内容は,短期的課題から長期的視点までを広く含有し,「AI をより有能にする研究だけでなく,社会的利益を最大化する研究にもフォーカス」した意欲的な取組みと言える.欧米において社会が望む AI システムへの取組みが始まった今,我々も進んでそれらに参加し,また一方で,日本やアジアにおいても,その利益や避けるべきリスクや望むことを検討し始める必要があると考えられる.

参考文献

[1] 総合科学技術 イノベーション会議 基本計画専門調査会,2015,第５期科学技術基本計画に向けた中間取りまとめ（案）平成 27 年 5 月 28 日.

[2] 江間有沙,2015,「人工知能と未来」プロジェクトから見る現在の課題,第 29 回人工知能学会全国大会論文集.

[3] Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter, Last updated January 23, 2015.
http://futureoflife.org/static/data/documents/research_priorities.pdf

[4] Autonomous Weapons: an Open Letter from AI & Robotics Researchers,2015.
http://futureoflife.org/AI/open_letter_autonomous_weapons

[5] 2015 Project Grants Recommended for Funding.
http://futureoflife.org/AI/2015awardees

[6] 本田康二郎,2013,工学倫理とロボット倫理,社会と倫理 第 28 号 p.21―36.