June 19th JAPS-DLMPST Symposium　5人の講演アブストラクト

"Transparency in AI: Identifying the Real Issue"
Takayuki Suzuki (The University of Tokyo)
アブストラクト 日本語版
人工知能は不透明である、そしてそのことが社会にさまざまなリスクをもたらすとしばしば言われる。しかし、人工知能がどのような意味で不透明であるかは、かならずしも明らかではない。多くの古典的なアルゴリズムは透明である。また、不透明な人工知能の典型と考えられている深層ニューラルネットワークの働きについても、さまざまな説明を与えることが可能である。さらに、人工知能以外の領域においては、われわれは透明性をそれほど重視しないこともある。
とはいえ、人工知能の働きが理解できない事例が存在することは間違いない。このことは、深層ニューラルネットワークにおいて、しばしば予期せぬ反応やバイアスを見出す理由ともなっている。それゆえ、人工知能に固有の透明性の問題が存在することは間違いないように思われる。それがどのような問題であるかを明確化するには、さらなる検討が必要である。人工知能と同様、人間の心の働きも、しばしば不透明であり、さまざまなバイアスや欠点をもっているが、人間の弱みや強みと人工知能の弱みや強みは相補的な関係にある。それゆえ、われわれは、人間の知能をコンピュータによって再現することよりもむしろ、人工知能を活用して人間のもつバイアスや欠点を克服することを目指すのが生産的であるように思われる。

Abstracts of the (first-round) talks…

Invited Talk 1
“Legitimacy, Authority, and the Political Value of Explanations”
Seth Lazar (Australian National University)
As rapid advances in Artificial Intelligence and the rise of some of history's most potent corporations meet the diminished neoliberal state, we have become increasingly subject to power exercised by means of automated systems. Machine learning, big data, and related computational technologies now underpin vital government services from criminal justice to tax auditing, from public health to social services, from immigration to defence. Two-sided markets connecting consumers and producers are shaped by algorithms proprietary to companies such as Google and Amazon. Google's search algorithm determines, for many of us, how we find out about everything from how to vote to where to get vaccinated; Facebook, Twitter and Google decide which of our fellow citizens' speech we get to see—both what gets taken down, and (more importantly) what gets promoted. We sometimes imagine AI as a far

off goal—either the handmaiden to a new post-scarcity world, or else humanity's apocalyptic 'final invention'. But we are already using AI to shape an increasing proportion of our online and offline lives. As the pandemic economic shock ramifies, and the role of technology in our lives grows exponentially, this will only intensify.

We are increasingly subject to Automatic Authorities—automated computational systems that are used to exercise power over us, by substantially determining what we may know, what we may have, and what our options will be. These computational systems promise radical efficiencies and new abilities. But, as is now widely recognised, they also pose new risks. In this paper I focus on one in particular: that the adoption of Automatic Authorities leads us to base increasingly important decisions on systems whose operations cannot be adequately explained to democratic citizens.

Philosophers have long debated the importance of justifications in morality and politics, but they have not done the same for explanations. What's more, the most prominent Automatic Authorities in our lives today are deployed by non-state actors like Google and Facebook, and analytical political philosophy has focused much more on state than non-state power. To make progress on one of the most pressing questions of the age of Automatic Authorities, therefore, we must make substantial first-order progress, on two fronts, in moral and political philosophy.

That is my goal in this paper. My central claim: only if the powerful can adequately explain their decisions to those on whose behalf or by whose licence they act can they exercise power legitimately and with proper authority, and so overcome presumptive objections to their exercise of power grounded in individual freedom, social equality, and collective self-determination. This applies to all authorities, not only automatic ones. But I will demonstrate its application to Automatic Authorities, including those sustained by non-state actors. I will use this account of why explanations matter to address the urgent regulatory questions of to whom explanations are owed, and what kinds of explanations are owed to them.

Invited Talk 2
“Explanation: from ethics to logic”
Gilles Dowek (INRIA and ENS Paris-Saclay)
Abstract: Explaining decisions is an ethical necessity. For example, neither a person nor a piece of software ought to reject a bank loan application, without providing an explanation for this rejection. But, defining the notion of explanation is a challenge. Starting from concrete examples, we attempt to understand what such a definition could look like.

A usual definition assumes that what is explained is a statement and what explains it is a logical proof of this statement. For example, a proof in elementary geometry both shows that the statement is true and explains why it is true. But this definition is not sufficient, as some logical proofs are seen as more explanatory than others.

In this talk, we give two successive definitions of the notion of explanation. The first keeps the idea that an explanation is a logical proof, but the explanatory character of the a proof relies on the fact that it contains a cut: the proof of general statements followed by a specialization to a particular one. So, the fact that a proof is explanatory is measured by the degree of generalization it allows. The second definition uses the algorithmic interpretation of proofs to generalize this definition. An explanation is then a pair formed with a short, fast, and wide algorithm and an input value for this algorithm.

Invited Talk 3
“Justified Representation in Approval-Based Committee Voting”
Edith Elkind (University of Oxford)
Computational social choice is a rapidly growing research field that studies algorithmic aspects of collective decision-making and preference aggregation. It studies questions such as: can we quickly compute the outcome of a given voting rule? For a given voting rule, can a strategic voter efficiently compute her optimal strategy? Is there a voting rule that satisfies a particular set of desirable properties (axioms) and admits an efficient winner determination algorithm?

In this talk, we consider these questions in the context of multiwinner voting rules with approval ballots. We formulate a fairness axiom, which we call Justified Representation, as well as a strengthening of this axiom. We identify voting rules that satisfy the JR axiom and investigate their algorithmic complexity. Finally, we propose a tractable rule that satisfies the strong version of the axiom.

Two Invited Discussant’s Talks
Invited Discussant’s Talk 1
“What are explanations worth in science?”
Minao Kukita (Nagoya University)
Steven Weinberg argues that the exemplar of physics is "a simple set of mathematical principles that govern a wide range of phenomena with precision. As can be seen here,

"generality," "precision," and "simplicity" are highly thought of in science. Then, why are they important in science? In this talk, I will discuss the value of generality and simplicity in particular from the perspective of the importance of information communication in science. Thus, we will focus on the value of explanations in science in reducing the cognitive cost of sharing and applying information.

I will argue that the increasing use of artificial intelligence based on big data in the practice of science means that information will no longer be shared and applied in the traditional way. Scientific explanations will then become irrelevant in terms of saving costs in information communication. However, scientific explanations also have aesthetic value beyond mere means of saving cognitive costs.How the perception of such aesthetic value will change in the future is an important topic, and I would like to call on people working in the fields of philosophy of science, sociology of science, and STS to explore this topic.

Invited Discussant's Talk 2
"Transparency in AI: Identifying the Real Issue"
Takayuki Suzuki (The University of Tokyo)
It is often argued that AI lacks transparency and that this might be a potential risk to our society. It is not clear, however, in what sense AI lacks transparency. Many classical algorithms are transparent. Even in deep neural networks, which are typically seen as opaque, we can have some explanation of how they work. Moreover, in areas other than AI, we often don't care much about transparency.

It is true, however, in some cases, how AI works is inscrutable. This is why we sometimes find unexpected responses and biases in AI, especially in deep neural networks. So, it seems that there really is a problem of transparency that is unique to AI. We need more work to identify what it is.

It also should be noted that human mind, as well as AI, is often opaque and biased. It will be more productive if we try to overcome our weaknesses with AI, rather than try to replicate our intelligence with AI, because the weaknesses and the strengths of human mind and those of AI complement each other.