Dagstuhl-Seminar

Ethics and Trust: Principles, Verification and Validation

22. – 26. April 2019

Pressemitteilung

Künstliche Intelligenz, Roboter, maschinelles Lernen und autonome Systeme spielen eine immer größere Rolle in unserer Gesellschaft. Technologien mit wachsender Bedeutung sind zum Beispiel:

- autonome, selbstgesteuerte Fahrzeuge;
- unbemannte, autonome Drohnen im Militär;
- Behandlungs- und Pflegeroboter sowie computergestützte Diagnosesysteme in der Medizin;
- autonome Systeme auf den Finanzmärkten, die eigenständig Investitionsentscheidungen treffen;
- maschinelle Systeme, die die Kreditwürdigkeit von Bankkunden beurteilen;
- Algorithmen, die über die Vergabe von Studienplätzen an Bewerber entscheiden;
- und sogar algorithmische Entscheidungssysteme in der Justiz (z. B. in den USA), die das Rückfälligkeitsrisiko von verurteilten Straftätern bewerten.

Mit der zunehmenden Anwendung solcher Systeme in der Gesellschaft wachsen die ethischen Herausforderungen. Wenn ein autonomes System mit einer schwierigen Entscheidungssituation konfrontiert wird, können wir uns darauf verlassen, dass das System die moralisch richtige Entscheidung trifft? Würde zum Beispiel ein selbstgesteuertes Fahrzeug eine fatale Kollision mit einem Fußgänger vermeiden, selbst wenn dies ein geringfügiges Verletzungsrisiko für die Fahrzeuginsassen beinhaltet? Würde eine autonome militärische Drohne Kollateralschäden vermeiden? Würde ein medizinischer Roboter die richtige Entscheidung in einer Notfallsituation treffen? Treffen algorithmische Entscheidungssysteme bei der Kreditvergabe oder erst recht in der Justiz möglicherweise diskriminierende Entscheidungen?

In all diesen Fällen stellen sich zwei Schlüsselfragen:

- Können sich künstlich-intelligente Systeme tatsächlich ethisch verhalten, und was genau bedeutet das?
- Können wir Menschen uns auf das ethische Verhalten solcher Systeme verlassen?

Mit diesen Fragen beschäftigt sich vom 22. bis zum 26. April ein internationales interdisziplinäres Seminar im Schloss Dagstuhl, dem Leibniz-Zentrum für Informatik. Die Teilnehmerinnen und Teilnehmer sind Wissenschaftler aus unterschiedlichen Ländern und aus den verschiedensten Fachbereichen: von der Informatik und der Mathematik über die Philosophie und Gesellschaftswissenschaften bis hin zu Wirtschafts- und Rechtswissenschaften. Auf Basis von Expertenvorträgen, fächerübergreifenden Diskussionen und Arbeitsgruppen beschäftigen sich die Teilnehmenden mit Fragen wie etwa:

- Ist es möglich, moralisches Entscheidungsverhalten in Computersystemen zu programmieren? Bedarf es einer grundsätzlichen Neuentwicklung unserer moralischen Konzepte und Prinzipien, um diese „computertauglich“ zu machen? Man bedenke, dass unsere traditionellen Konzepte der Moral ausschließlich für die Verwendung durch den Menschen entwickelt worden sind.
- Und wie kann sich die Gesellschaft davon überzeugen, dass die entsprechenden Systeme nicht nur aus technischer, sondern auch aus moralischer Sicht zuverlässig sind? Viele Menschen sind der Überzeugung, dass die Politik und die öffentliche Verwaltung die Verwendung autonomer Systeme in risikobehafteten Bereichen erst dann zulassen sollten, wenn die technische und ethische Sicherheit solcher Systeme hinreichend geprüft ist. Welche technischen, ethischen und gesellschaftlichen Maßnahmen sind notwendig, um die Verwendung autonomer Systeme in riskanten Anwendungsbereichen angemessen zu regulieren?

Um diese und ähnliche Fragen zu beantworten, ist eine sorgfältige Grundlagenforschung erforderlich, und es bedarf einer engen Zusammenarbeit von theoretischen Forschern aus der Philosophie, Ethik und Rechtswissenschaft, technischen Forschern aus der Informatik, Robotik und Ingenieurwissenschaft und gesellschaftswissenschaftlichen Forschern aus der Wirtschaft, Politik und Soziologie. Selbst grundsätzliche Fragen – wie etwa, ob die Moral ein und für alle Male in expliziten Prinzipien zusammengefasst werden kann oder ob die Moral ein flexibles, sich ständig weiterentwickelndes System sein sollte – gilt es zu klären.

Zu dieser Thematik ist sehr viel wissenschaftliches Neuland zu erschließen, und Ziel des einwöchigen Seminars im Schloss Dagstuhl ist es, durch interdisziplinäre Zusammenarbeit Fortschritte zu machen.

Englische Seminarbeschreibung

Artificial morality, also called machine ethics, is an emerging field in artificial intelligence that explores how autonomous systems can be enhanced with sensitivity and respect for the legal, social, and ethical norms of human society. Academics, engineers, and the public at large, are all wary of autonomous systems, particularly robots, drones, “driverless” cars, etc. Robots will share our physical space, and so how will this change us? With the predictions in hand of roboticists we can paint portraits of how these technical advances will lead to new experiences and how these experiences may change the ways we function in society. Two key issues are dominant, once robot technologies have advanced and yielded new ways we and robots share the world:

1. will robots behave ethically, i.e.: as we would want them to, and
2. can we trust them to act to our benefit.

Rather than any engineering issues, it is these barriers concerning ethics and trust that are holding back the development and use of autonomous systems. One of the hardest challenges in robotics seems to be reliably determining desirable and undesirable behaviours for robots. Our aim here is to advance the work in these areas, bringing together a range of disciplines that can impact upon these problems.

Some of us organised the Dagstuhl 16222 Engineering Moral agents: From human morality to artificial morality seminar in 2016 with the goal of initiating a conversation between Philosophers studying ethics, Robotics researchers developing novel autonomous machines, and Computer Scientists studying AI and reasoning. This provides a clearer understanding of the issues and several avenues for future collaboration. However, it also highlighted further important areas to be exposed, specifically:

- the extension of 'ethics' to also address issues of 'trust'
- the practical problems of implementing ethical and trustworthy autonomous machines; and
- the new verification and validation techniques that will be required to assess these dimensions.

We expect the seminar to:

- Give researchers across the contributing disciplines an integrated overview of current research in machine ethics and trustworthy robotics from the artificial intelligence side and of relevant areas of philosophy and psychology.
- Open up a communication channel among researchers tackling ethics and trust, bridging the computer science/humanities/social-science divide in these fields.
- Identify the central research questions and challenges concerning (i) the definition and operationalisation of the concept of ethics and trust in autonomous systems; (ii) the formalisation and algorithmization of theories of ethics and trust; and (iii) the relationships between ethics and trust in both human and non-human systems.
- Identify existing and potential societal consequences of these systems. What are the risks, what are the chances, what are beneficial use cases for these systems?

Artificial ethics and trust between humans and autonomous entities both bring together many disciplines which have a vast amount of relevant knowledge and expertise, but which are often inaccessible to one another and insufficiently develop their mutual synergies. Researchers need to communicate to each other their experiences, research interests, and knowledge to move forward. We plan the seminar as a combination of three structures: tutorials, contributed talks, and discussion sessions.

Organisatoren

Michael Fisher (University of Liverpool, Großbritannien)
Christian List (London School of Economics, Großbritannien)
Marija Slavkovik (University of Bergen, Norwegen)
Astrid Weiss (TU Wien, Österreich)