– For German version, see below –

Stiftung Mercator awards €3.8 Million to the Universities of Bonn and Cambridge to Research Ethical AI

Artificial intelligence (AI) is transforming society as algorithms increasingly impact access to jobs and insurance, justice, medical treatments, as well as our daily interactions with friends and family. As these technologies race ahead, we are starting to see unintended social consequences: algorithms that promote everything from racial bias in healthcare to the misinformation eroding faith in democracies.

To ensure AI supports core human values, the German philanthropic foundation Stiftung Mercator has awarded a €3.8 million grant to a collaboration between the Universities of Bonn and Cambridge. Led by Professor Markus Gabriel from the Institute for Philosophy at Bonn and Dr Stephen Cave from the Leverhulme Centre for the Future of Intelligence at Cambridge, the project, '*Desirable Digitalisation: Rethinking AI for Just and Sustainable Futures*', places ethical principles at the heart of AI development.

The new research project comes as the European Commission negotiates its Artificial Intelligence Act, which has ambitions to ensure AI becomes more "trustworthy" and "human-centric". The Act will require AI systems to be assessed for their impact on fundamental rights and values. The researchers on the Desirable Digitalisation project will collaboratively investigate the many questions that arise from these plans, such as: What exactly does a "human-centric" approach to AI look like? How can we meaningfully assess whether and how AI systems violate fundamental rights and values? And how can we foster awareness of discriminatory practices and how to stop them?

Carla Hustedt, director of Stiftung Mercator's Centre for Digital Society, explains: "The socio-technological nature of AI systems requires us to break out of silos in multiple ways: We need interdisciplinary, international research as well as the cooperation between scientific actors with actors from business and civil society. The project seeks to do exactly that."

The Desirable Digitalisation project is divided into two parts. In the first part, researchers will investigate intercultural perspectives on AI and fundamental rights and values. As Dr Cave explains: "In order to understand the potential effects of algorithms on human dignity, we need to look beyond the code and draw on lessons from history and political science." This part of the project will ask questions from two perspectives: anthropological (How will our idea of 'the human' influence and be influenced by digital technology?) and intersectional (How do the structural injustices of the past influence today's technology and its influence on fundamental rights and values?).

The Cambridge and Bonn teams will work not only with colleagues across Europe, but also with teams in Asia and Africa. As Professor Gabriel points out: “Irrespective of our specific cultural world-views, these new technologies challenge our idea of ourselves as human beings.” The project therefore investigates foundational, anthropological questions concerning the human in the digital age. How do different ideas of the human shape different cultures’ views of desirable digitalization?

In the second part of the project, ‘Designing AI for Just and Sustainable Futures’, researchers from both universities will work with the AI industry to develop design and education principles that put sustainability and justice at the heart of technological progress. According to Prof. Aimee van Wynsberghe, Humboldt Professor at the University of Bonn, who will lead the Bonn team in this second part of the project and will be contributing her expertise on sustainability: “Sustainability in all its dimensions – social, ecological, economic and technological – is a vital value in designing AI. Only by taking it into account can these technologies improve our lives and our world.”

The five-year project will start in April 2022, with the first of its biannual conferences taking place in early 2023. The core team of seventeen researchers at Cambridge and Bonn, as well as visiting professors, will work closely with a wide range of national and international partners.

For more information, contact:

University of Bonn:
Jan Voosholz
Institut für Philosophie
Universität Bonn
Email: voosholz@uni-bonn.de

University of Cambridge:
Dr Kanta Dihal
Leverhulme Centre for the Future of Intelligence
University of Cambridge
Email: ksd38@cam.ac.uk

Stiftung Mercator:
David Alders
Centre for Digital Society
Stiftung Mercator
Email: david.alders@stiftung-mercator.de

Stiftung Mercator vergibt 3.8 Millionen Euro für die Universitäten Bonn und Cambridge zur Erforschung von ethischer KI

Künstliche Intelligenz (KI) verändert die Gesellschaft tiefgreifend: Algorithmen beeinflussen den Zugang zu Berufen und Versicherungen, zur Justiz und medizinischen Behandlungen sowie unsere alltäglichen Interaktionen mit Freund*innen und Familie. Je schneller sich diese Technologien entwickeln, desto deutlicher werden die gesellschaftlichen Folgen ihres Einsatzes: Algorithmen, die rassistische Vorurteile im Gesundheitswesen fördern bis hin zur Verbreitung von Falschinformationen, die das Vertrauen in Demokratien untergraben.

Um einen Beitrag zu leisten, dass KI grundlegende menschliche Werte unterstützt, fördert die Stiftung Mercator die Zusammenarbeit zwischen den Universitäten Bonn und Cambridge mit 3,8 Millionen Euro. Unter Leitung von Prof. Markus Gabriel des Bonner Instituts für Philosophie und Dr. Stephen Cave vom Leverhulme Centre for the Future of Intelligence in Cambridge wird das Projekt *Wünschenswerte Digitalisierung (Desirable Digitalisation: Rethinking AI for Just and Sustainable Futures*) ethische Prinzipien in den Mittelpunkt der KI-Entwicklung stellen.

Das Projekt startet während auf europäischer Ebene das *Gesetz über Künstliche Intelligenz* verhandelt wird, das einen Rechtsrahmen für „vertrauenswürdige“ KI zu schaffen versucht, die „auf den Menschen ausgerichtet“ ist. Dieses Gesetz wird verlangen, dass der Einfluss von KI-Systemen auf grundlegende Rechte und Werte geprüft und bewertet wird. Wie genau kann KI „auf den Menschen ausgerichtet“ sein? Wie können wir sinnvoll beurteilen, ob und wie KI-Systeme grundlegende Rechte und Werte verletzen? Und wie können wir das Bewusstsein für diskriminierende Praktiken stärken und diese stoppen?

Carla Hustedt, Leiterin des Bereichs Digitalisierte Gesellschaft der Stiftung Mercator erklärt: „KI-Systeme sind soziotechnische Systeme. Ihre Gestaltung verlangt von uns, dass wir verschiedene Silos aufbrechen: Wir brauchen interdisziplinäre, internationale Forschung sowie die Zusammenarbeit von wissenschaftlichen Akteuren mit Akteuren aus Wirtschaft und Zivilgesellschaft. Genau hier setzt das Projekt an.“

Das Projekt *Wünschenswerte Digitalisierung* ist in zwei Teile geteilt. Im ersten Teil untersuchen Forscher*innen interkulturelle Perspektiven auf KI und grundlegende Rechte und Werte. So macht Dr. Cave deutlich: „Um die potenziellen Auswirkungen von Algorithmen auf die Menschenwürde zu verstehen, müssen wir über den Code hinausblicken und Lehren aus Geschichte und Politikwissenschaft ziehen.“ In diesem Teil des Projekts werden Fragen aus zwei Blickwinkeln gestellt: aus anthropologischer Sicht (Wie wird unsere Vorstellung vom „Menschen“ durch digitale Technologien beeinflusst und beeinflusst umgekehrt diese?) und aus intersektionaler Sicht (Wie prägen strukturelle und historische Ungerechtigkeiten die Entwicklung und den Einsatz von Technologie? Wie wirkt sich das auf Grundrechte und Werte aus?)

Die Teams in Cambridge und Bonn werden sowohl mit Kolleg*innen aus Europa als auch mit Teams in Asien und Afrika zusammenarbeiten. So betont Prof. Gabriel: „Unabhängig von unseren spezifischen kulturellen Weltanschauungen stellen diese neuen Technologien unser Selbstverständnis als menschliche Wesen in Frage.“ Daher beschäftigt sich das Projekt mit grundlegenden anthropologischen Fragen über den Menschen im digitalen Zeitalter. Wie

prägen unterschiedliche Ideen über den Menschen die Ansichten verschiedener Kulturen über wünschenswerte Digitalisierung?

Im zweiten Teil des Projekts *Designing AI for Just and Sustainable Futures*, werden Forscher*innen beider Universitäten mit Partner*innen in der KI-Industrie zusammenarbeiten, um Prinzipien für das Design und die Bildung von KI zu entwickeln, die Nachhaltigkeit und Gerechtigkeit in den Mittelpunkt des technologischen Fortschritts stellen. Prof. Aimee van Wynsberghe, Humboldt Professorin an der Universität Bonn, die das Bonner Team im zweiten Teil leiten und ihre Expertise zu Nachhaltigkeit einbringen wird, erklärt: „Nachhaltigkeit in all ihren Dimensionen – sozial, ökologisch, wirtschaftlich und technologisch – ist ein wesentlicher Wert bei der Entwicklung von KI. Nur wenn sie berücksichtigt wird, können diese Technologien unser Leben und unsere Welt verbessern."

Das Projekt startet im April 2022 mit einer Laufzeit von fünf Jahren; die erste der zweijährlichen Konferenzen findet Anfang 2023 statt. Das Kernteam, bestehend aus siebzehn Forscher*innen aus Cambridge und Bonn und Gastprofessor*innen, wird mit einer Vielzahl nationaler und internationaler Partner eng zusammenarbeiten.

Für weitere Informationen kontaktieren Sie:

Universität Bonn:
Jan Voosholz
Institut für Philosophie
Universität Bonn
E-Mail: voosholz@uni-bonn.de

Universität Cambridge:
Dr. Kanta Dihal
Leverhulme Centre for the Future of Intelligence
University of Cambridge
E-Mail: ksd38@cam.ac.uk

Stiftung Mercator:
David Alders
Stiftung Mercator
Bereich Digitalisierte Gesellschaft
E-Mail: david.alders@stiftung-mercator.de