Handreichung zum Thema Ethik am DFKI

1. Über dieses Dokument

Diese Handreichung dient dem Zweck der Einführung des Ethik-Teams am DFKI. Um unseren Arbeitsbereich zu beschreiben, sprechen wir eine Reihe von grundlegenden Themen an, die relevant für einen Diskurs über KI und Ethik sind. Konkrete Maßnahmen, Richtlinien oder gar Handlungsempfehlungen sind hier (noch) nicht zu erwarten. Diese setzen eine intensive Auseinandersetzung mit konkreten Fällen voraus, die erst in Zukunft erfolgen wird. Aus dieser Auseinandersetzung hoffen wir verallgemeinerbare Aspekte in zukünftige Versionen dieser Handreichung einfließen lassen zu können.

2. Einführung und Motivation

Während menschliche Individuen ihr Handeln an moralischen Vorstellungen messen, reflektiert die Ethik das moralische Handeln. Über die Frage, ob man bei KI-Systemen überhaupt von (ggf. autonomem) Handeln sprechen kann, gibt es keinen Konsens. Sicher ist aber, dass KI-Systeme aus menschlichen Handlungen lernen können und dass (autonome) KI-Systeme das Leben von Individuen und die Gesellschaft als Ganzes beeinflussen können, etwa, indem Entscheidungen unterstützt werden oder die Systeme physisch mit Menschen interagieren. Aus diesem Grund ist es notwendig, bei der Forschung und Entwicklung von KI-Systemen ethische Aspekte zu betrachten, wobei der Umfang der Betrachtung je nach Forschungsvorhaben sehr stark variiert.

3. Ziel, Umfang, Modus

Das Ziel dieser kurzen Handreichung ist es, DFKI-externe Personen über diese Policy des DFKI zu informieren. Es ist aus unserer Sicht nicht möglich, detaillierte ethische Erfordernisse für die Arbeiten in mehr als 20 Forschungsgruppen und hunderten völlig verschiedener Projekte zu formulieren. Daher wollen wir im Folgenden nur einige Impulse setzen.

4. Ethische Grundsätze für die Forschung und Entwicklung am DFKI

Für die Forschung und Entwicklung an KI-Systemen am DFKI sind die folgenden Grundsätze maßgeblich.

KI zum Wohle des Menschen

Bei allem Denken und Tun innerhalb des DFKI steht das Wohl des Menschen als Individuum und der Menschheit als Ganzes im Mittelpunkt. Als Basis dienen die Menschenrechte; wir betrachten den Menschen jederzeit als Subjekt, niemals als Objekt. Dies wird u.a. auch durch den Leitspruch des DFKI „KI für den Menschen“ zum Ausdruck gebracht.

KI und Nachhaltigkeit
Das DFKI strebt mit seiner Forschung und Entwicklung an, die globalen Nachhaltigkeitsziele der UN [Bundesregierung 2019] aktiv voranzubringen.

Das Prinzip der Akzeptabilität
Für jegliche Interaktion eines Menschen mit einem KI-System soll das Prinzip der Akzeptabilität [Alpsancar 2018] gelten: Menschen müssen in die Lage versetzt werden, eine zustimmende oder ablehnende Position gegenüber dem KI-System einzunehmen. Hierfür muss für den menschlichen Nutzer jederzeit transparent sein, ob er sich in einer Interaktionssituation mit dem KI-System befindet oder nicht. Ebenfalls muss klar sein, wann und wie er diese Interaktionssituation verlassen kann, ggf. unter Nennung möglicher Konsequenzen dieser Entscheidung. Desweiteren müssen die Abläufe innerhalb des Systems dem Menschen bei Bedarf erklärt werden können. Das Prinzip der Akzeptabilität soll der Gestaltung jedes am DFKI entwickelten Systems zugrunde liegen.

Transparenz schaffen auch über die Grenzen der Systeme
Um die Zuverlässigkeit und Nutzbarkeit eines KI-Systems in jeder Hinsicht adäquat bewerten zu können ist es erforderlich, auch seine Grenzen zu kennen. Daher soll für jedes am DFKI entwickelte System in hinreichender Granularität dokumentiert werden, für welche Aufgabenstellungen es geeignet ist und für welche nicht. Bei einem neuronalen Netz soll beispielsweise angegeben werden, mit welcher Art Daten es trainiert wurde (z.B. „Dieses System wurde ausschließlich mit Fotos von Europäern trainiert.“). So kann ein Nutzer verstehen, welche Ausgaben bei bestimmten Eingaben zu erwarten sind bzw. ob für bestimmte Eingaben überhaupt sinnvolle Ausgaben erwartbar sind. Wo möglich, sollen gerade potenzielle Randfälle in den Test solcher Systeme einfließen.

KI und Bewusstsein
Das DFKI konzentriert sich in seinen Forschungen auf das Feld der schwachen KI, die vom Menschen im Sinne eines Werkzeugs verwendet werden kann. Das DFKI unterstützt keine Forschung, die darauf abzielt, eine künstliche Intelligenz mit Bewusstsein oder mit eigenen komplexen Zielen zu erschaffen.

Die Forschung an KI für militärische Zwecke wird eingeschränkt
Das DFKI beteiligt sich an Forschung und Entwicklung von KI für militärische Zwecke in einem eng gesteckten Rahmen, der in einem separaten internen Dokument beschrieben wird. Hierzu gibt es einen entsprechenden Beschluss des Lenkungskreises.

5. Unterstützung und Mitarbeit
Das Ethik-Team des DFKI wird durch den Lenkungskreis des DFKI eingesetzt und durch die Geschäftsleitung legitimiert. Es besteht zurzeit aus Aljoscha Burchardt, Christiane Plociennik

und Christian Müller und wird unterstützt durch Antonio Krüger, Gesche Joost und Paul Lukowicz. Es steht der Belegschaft als Ansprechpartner in allen Ethik-Fragen zur Verfügung, sei es bezogen auf ein konkretes Projekt oder allgemein die Arbeit am DFKI betreffend. Am DFKI kann auf Erfahrung bei der Einbindung von externen Ethik-Experten in Projekte mit entsprechendem Bedarf zurückgegriffen werden. Alle Mitarbeiterinnen und Mitarbeiter des DFKI sind eingeladen, die Handreichung gemeinsam mit dem Ethik-Team weiterzuentwickeln.

6. Referenzen

[Alpsancar 2018] Alpsancar, Suzana (2018): Die Ethik der Künstlichen Intelligenz. https://www-docs.b-tu.de/fg-technikwissenschaft/public/BTU_News_09_2018_Die_Ethik_der_k%C3%BCnstlichen_Intelligenz.pdf, abgerufen am 16.01.2020.

[Bundesregierung 2019] Die Bundesregierung: Nachhaltigkeitsziele verständlich erklärt. https://www.bundesregierung.de/breg-de/themen/nachhaltigkeitspolitik/nachhaltigkeitsziele-verstaendlich-erklaert-232174, abgerufen am 16.01.2020.