

# ЕТИЧНИ НАСОКИ ЗА НАДЕЖДЕН AI

От Групата от Експерти на  
Високо Ниво върху  
Изкуствения Интелект

Giovanni Sartor



# Документът

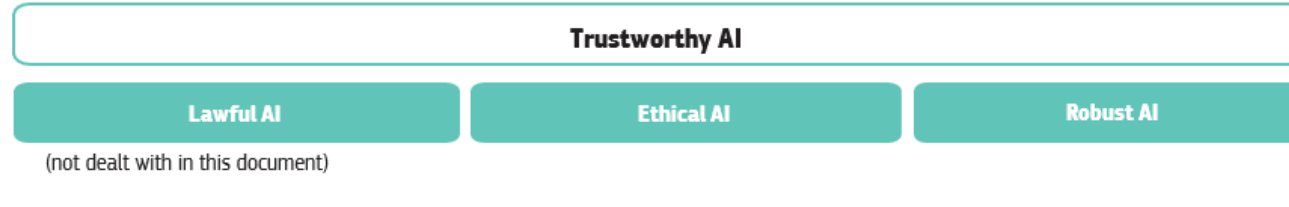
- Изготвен от Групата от Експерти на Високо ниво върху Изкуствения Интелект от Европейската комисия през юни 2018 г.
- публикуван на 8 април 2019 г.
- достъпно онлайн (<https://ec.europa.eu/digital-single-market/en/high-livello-esperto-gruppo-artificiale-intelligenza>).
- Това е добър пример за многото публикувани документи за етиката на AI до тук

# Идеята за надежден AI

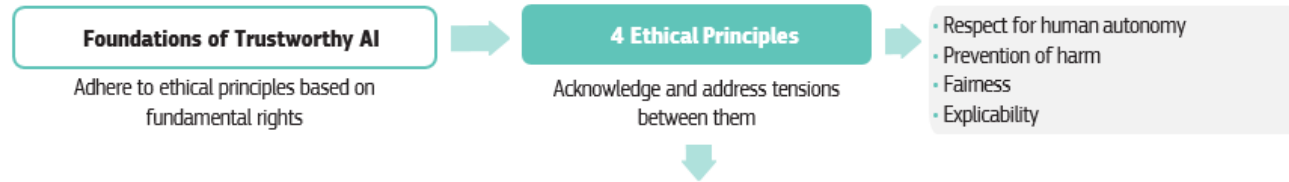
- AI трябва да бъде
  - Законен, спазващ всички приложими закони и разпоредби
  - Етичен, гарантиращ придържане към етичните принципи и ценности
  - Здрав, както от техническа, така и от социална гледна точка, тъй като, дори и с добри намерения, AI системите могат да причинят неволна вреда
- Тези изисквания трябва да бъдат изпълнени през целия животен цикъл на системата
- Въпрос. Сещате ли се за примери за незаконни, неетични или не-надеждни употреби на AI?

# Framework for Trustworthy AI

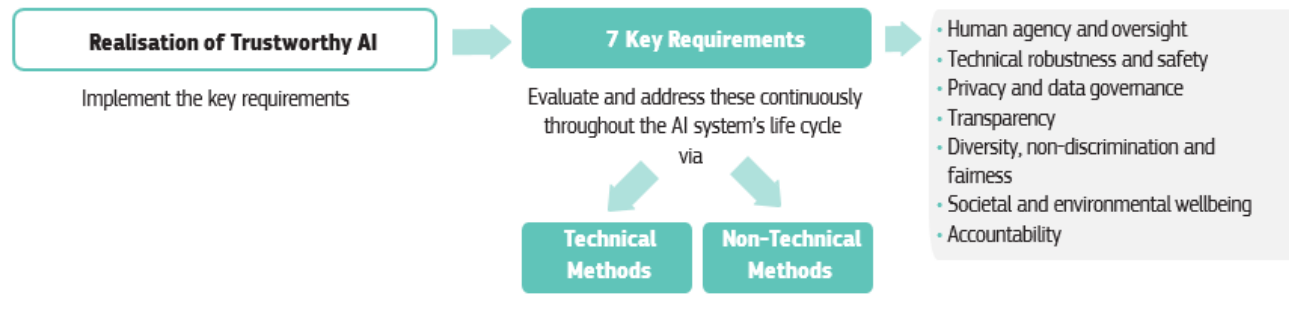
INTRODUCTION



CHAPTER I



CHAPTER II



CHAPTER III



# Глава 1: Етични принципи

Разработвайте, внедрявайте и използвайте AI системи по начин, който се придържа към етичния принцип:

- уважение към човешката автономия,
  - предотвратяване на вреди,
  - справедливост и
  - обяснимост.
- Признайте и обърнете внимание на потенциалните напрежения между тези принципи.
  - Обърнете особено внимание на
    - ситуации, включващи по-уязвими групи като деца, хора с увреждания и други, които имат исторически са били в неравностойно положение или са изложени на риск от изключване, и
    - ситуации, които се характеризират с асиметрия на власт или информация, като например между работодатели и работници или между бизнеса и потребителите.
  - Признайте, че макар да носят значителни ползи за хората и обществото,
    - Системите с изкуствен интелект също представляват определени рискове и могат да имат отрицателно въздействие, включително въздействия, които може да са трудни за отстраняване предвиждат, идентифицират или измерват (напр. относно демокрацията, върховенството на закона и разпределителната справедливост, или върху човешкия самия ум.)
    - Приемете адекватни мерки за смекчаване на тези рискове, когато е подходящо и пропорционално на големината на риска.



# Глава II: насоки за реализация на надежден AI

- Гарантиране, че разработването, внедряването и използването на AI системи отговаря на седемте ключови изисквания за надежден AI:
  - (1) човешка агенция и надзор,
  - (2) техническа устойчивост и безопасност,
  - (3) поверителност и управление на данните,
  - (4) прозрачност,
  - (5) разнообразие, недискриминация и справедливост,
  - (6) благосъстоянието на околната среда и обществото и
  - (7) отчетност.
- Обмислете технически и нетехнически методи, за да гарантирате изпълнението на тези изисквания.

# Глава II: насоки за реализация на надежден AI (продължение)

- Насърчаване на изследванията и иновациите
  - да подпомогне оценката на системите за изкуствен интелект и да продължи постигането на изискванията; разпространяват резултатите и отваряне на въпроси към широката общественост и систематично обучение на ново поколение експерти в AI етика.
- Комуничайте по ясен и проактивен начин информация на заинтересованите страни относно AI възможностите и ограниченията на системата,
  - създаване на възможност за реалистични очаквания и за начина, по който са изискванията изпълнени. Бъдете прозрачни относно факта, че имат работа с AI система.
- Улесняване на проследимостта и възможността за одит на AI системите
  - особено в критични контексти или ситуации.
- Включете заинтересованите страни през целия жизнен цикъл на AI системата.
  - Насърчаване на обучението и образованието, така че всички заинтересовани страни да са запознати и обучени в Trustworthy AI.
- Имайте предвид, че може да има фундаментално напрежение между различни принципи и изисквания.
  - Непрекъснато идентифицирайте, оценявайте, документирайте и съобщавайте тези компромиси и техните решения.

# Глава III: Оценяване на надеждния AI

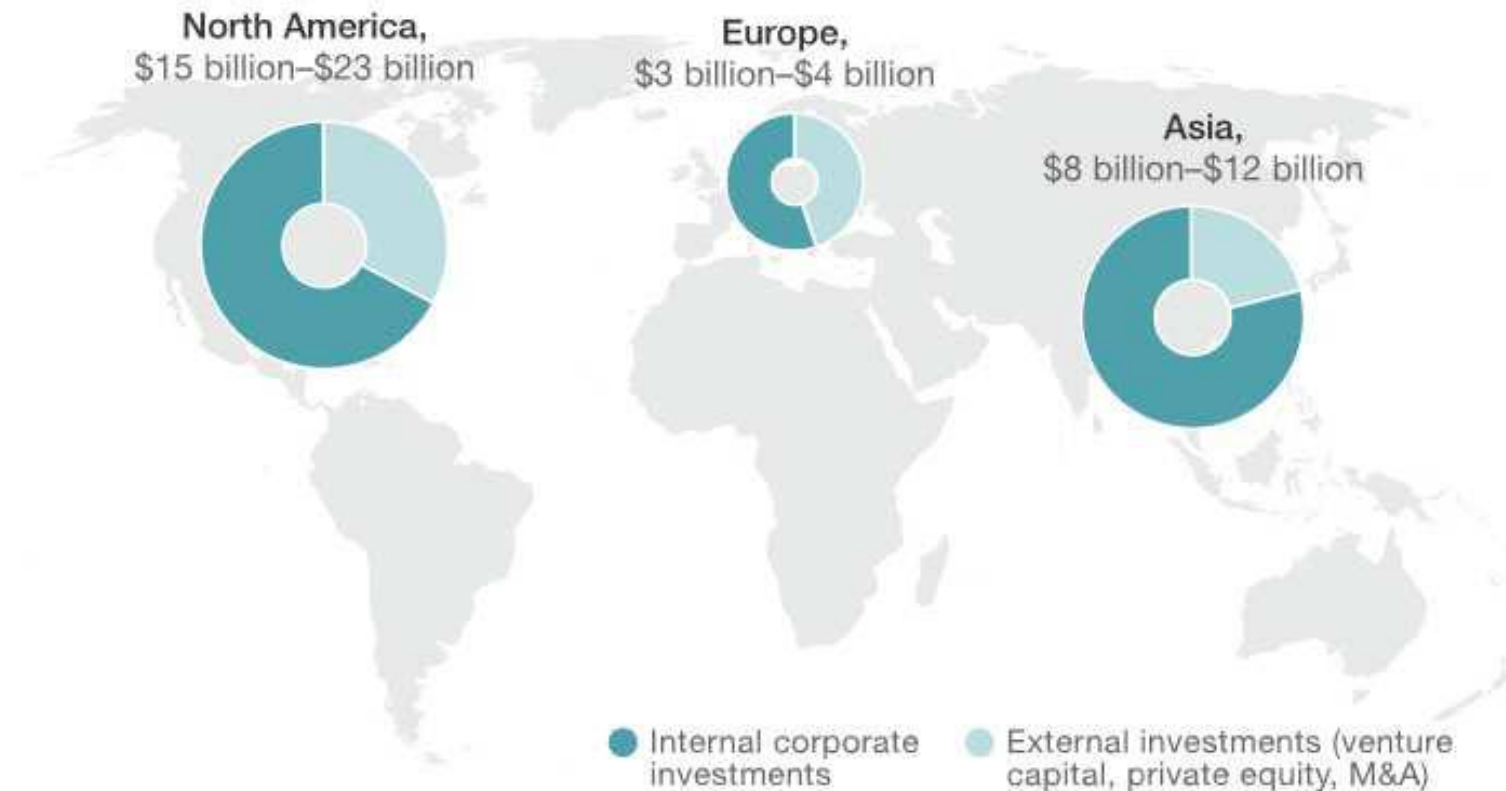
- Приемете списък за оценяване на надеждния AI
  - при разработване, внедряване или използване на AI системи и адаптирането им към конкретните условия в случая на употреба, в който се прилага системата.
- Имайте предвид, че такъв списък за оценка никога няма да бъде изчерпателен.
  - Осигуряването на надежден AI не е свързано с поставяне на отметки, а с непрекъснато идентифициране и прилагане на изисквания, оценка на решения, осигуряване на подобрени резултати през целия жизнен цикъл на AI системата и включващи заинтересованите страни в това.



# Подходът на Комисията към AI

- Съобщения от 25 април 2018 г. и 7 декември 2018 г. (COM(2018)237 и COM(2018)795). Три части:
  - (i) увеличаване на публичните и частните инвестиции в AI, за да се стимулира неговото усвояване
  - (ii) подготовка за социално-икономически промени, и
  - (iii) осигуряване на подходяща етична и правна рамка за укрепване на европейски ценности.
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF>
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-795-F1-EN-MAIN-PART-1.PDF>

# Проблем: наистина ли можем да се съпоставим със САЩ и Китай?



- <https://ec.europa.eu/growth/tools-databases/dem/monitor/content/usa-china-eu-plans-ai-where-do-we-stand>

# AI, ориентиран към човека

- ангажираност с използването на AI в услуга на човечеството и на общо благо, с цел подобряване на човешкото благосъстояние и свобода.
- увеличаване максимално предимствата на AI системите, като в същото време предотвратяване и минимизиране на рисковете от тях.

# Етика срещу закон

- Етика: норми, посочващи какво трябва да се прави по отношение на всички заложен интереси
  - Позитивна етика: норми, споделени в обществото (евентуално включващи идеи за социална йерархия, роли на половете и др.)
  - Критична етика: норми, които се разглеждат като най-подходящи или рационални
- Закон: норми, приети чрез институционални процеси и принудително наложени.

# Насоките за надежден AI като а (критична) етика?

- Заинтересованите страни, ангажирани с постигането на надежден AI, могат доброволно да изберат да използват тези насоки като метод за оперативно изпълнение на своя ангажимент,
- Насоките са адресирани до всички заинтересовани страни в AI, които проектират, разработват, внедряват, използват или се влияят от AI,
  - включително, но не само компании, организации, изследователи, обществени услуги, държавни агенции, институции, организации на гражданското общество, лица, работници и потребители.
- "Нищо в този документ не създава законови права, нито налага правни задължения спрямо трети страни. Припомняме обаче, че това е задължение на всички физически или юридически лица, да спазват законите – независимо дали са приложими днес или приети в бъдеще според развитието на AI."
- Каква е ролята на етиката спрямо закона в областта на AI?

# AI трябва да е законен

- Трябва да отговаря на
  - Първичния закон на ЕС (Договорите на Европейския съюз и неговата Харта за основните права),
  - Вторичния закон на ЕС (регламенти и директиви, като например Регламента за Общата защита на данните, Директивата относно продуктовата отговорност, Регламента за свободния поток на не-Лични данни, антидискриминационни директиви, потребителско право и безопасност и здраве при работни директиви),
  - Договорите на ООН за правата на човека и конвенциите на Съвета на Европа (като Европейска конвенция за правата на човека),
  - Закони на държавите-членки на ЕС (италианско право).
- Законите могат да бъдат хоризонтални на специфични за домейна правила (напр. за медицински устройства)
- Проблем: Можете ли да се сетите за хоризонтален закон, обхващащ всички приложения на AI?

# Основи на надежден AI

Етиката на AI е подполе на приложната етика,

- фокусиране върху етичните въпроси повдигнати от развитието, внедряване и използване на AI.
- Основната му грижа е да идентифицира как AI може да напредне или да повдигне въпроси за добрия живот на индивидите, независимо дали по отношение на качеството на живот, или човешка автономия и свобода, необходима за едно демократично общество.



# Основа: (етични) основни права

- Зачитане на човешкото достойнство. Човешкото достойнство обхваща идеята, че всяко човешко същество притежава „вътрешна стойност“
- Свобода на личността. Човешките същества трябва да останат свободни да създават житейски решения за себе си: включително (наред с други права) защита на свободата на стопанска дейност, свободата на изкуство и наука, свобода на изразяване, право на личен живот и неприкосновеността на личния живот и свободата на събиране и сдружаване.



# Основа: (етични) основни права

- Зачитане на демокрацията, справедливостта и върховенството на закона. AI системите трябва да не подкопават демократичните процеси, човешкото обсъждане или демократичните системи за гласуване, справедливия процес и равенството пред закона
- Равенство, недискриминация и солидарност – включително правата на лица в риск от изключване. В контекста на AI равенството предполага, че операциите на системата не могат да генерират несправедливо предубедени резултати. (GS: ние трябва да разберете какво означава това)
- Права на други граждани на право на глас, право на добра администрация или достъп до публични документи и правото на петиция до администрацията

# Етични принципи (основаващи се на правата на човека)

- (i) Зачитане на човешката автономия
- (II) Предотвратяване на вреди
- (III) Справедливост
- (IV) Обяснимост

# Уважение към човешката автономия

- Хората, взаимодействащи със системите за AI, трябва да могат да поддържат пълно и ефективно самоопределение над себе си и да могат да участват в демократичния процес.
  - AI системите не трябва неоправдано да подчиняват, принуждават, мамят, манипулират, да състояват или да насочват хора.
  - те трябва да бъдат създадени да увеличават, допълват и подобряват човешките когнитивни, социални и културни умения.
  - Разпределението на функциите между хората и AI системите трябва да следва принципите на дизайна, ориентирани към човека, и да оставя значима възможност за човешкия избор.
  - Това означава осигуряване на човешки надзор върху работните процеси в AI системите, подпомагане на хората в работната среда и стремеж към създаване на смислена работа.

# Принципът на предотвратяване на вредите

- Системите с AI не трябва нито да причиняват, нито да влошават дадена вреда или по някакъв друг начин да влияят неблагоприятно на хората.
  - Това включва защита на човешкото достойнство, както и психическата и физическа неприкосновеност.
  - AI системите и средите, в които работят, трябва да бъдат безопасни и сигурни.

# Принципът на справедливостта

- Съдържателно измерение
  - осигуряване на равно и справедливо разпределение както на ползите, така и на разходите, и
  - гарантиране, че лицата и групите са свободни от несправедливи пристрастия, дискриминация и стигматизация.
  - Насърчаване на равните възможности по отношение на достъпа до образование, стоки, услуги и технологии.
  - Никога не води до измама или неоснователно накърняване на свободата им на избор.
  - Практиците в областта на AI трябва да спазват принципа на пропорционалност между средствата и целите и да обмислят внимателно как да балансират конкуриращи се интереси и цели
- Процедурно измерение.
  - способност за оспорване и търсене на ефективна компенсация срещу решенията, взети от системите с AI и от хората, които ги управляват
    - За да се направи това, субектът, отговорен за решението, трябва да може да се идентифицира, а процесите на вземане на решение трябва да бъдат обясними.

# Принципът на обяснимостта

- За осигуряване на оспорваемост.
  - процесите трябва да са прозрачни,
  - възможностите и предназначението на системите за изкуствен интелект се съобщават открито, и
  - решения, доколкото е възможно, обясними на пряко и косвено засегнатите.
- Обяснение защо даден модел е генерирал конкретен резултат или решение (и каква комбинация от входни фактори е допринесла за това) не винаги е възможно.
  - може да са необходими други мерки за обяснение (напр. възможност за проследяване, възможност за одит и прозрачна комуникация относно възможностите на системата), при условие че системата като цяло зачита основните права.
  - степента, до която е необходима обяснимост, зависи силно от контекста и тежестта на последствията, ако този резултат е грешен или по друг начин неточен.

# Напрежение между принципите

- Трябва да се установят методи за отговорно обсъждане за справяне с такова напрежение.
- Конфликти между превенцията на вредата и човешката автономия
- Също така между благосъстоянието и сигурността?

# Изисквания за надежден AI

- 1. Човешка агенция и надзор
  - Включително основни права, човешка агенция и човешки надзор
- 2. Техническа издръжливост и безопасност
  - Включително устойчивост на атаки и сигурност, резервен план и обща безопасност, точност, надеждност и възпроизводимост
- 3. Поверителност и управление на данните
  - Включително уважение към неприкосновеността на личния живот, качеството и целостта на данните и достъпа до данните
- 4. Прозрачност
  - Включително проследимост, обяснимост и комуникация



# Изисквания за надежден AI (продължение)

- 5. Разнообразие, недискриминация и справедливост
  - Включително избягване на несправедливи пристрастия, достъпност и универсален дизайн и участие на заинтересованите страни
- 6. Обществено и екологично благополучие
  - Включително устойчивост и екологосъобразност, социално въздействие, общество и демокрация
- 7. Отговорност
  - Включително възможност за проверка, минимизиране и докладване на отрицателното въздействие, компромиси и обезщетение.



# Човешка агенция и надзор

- AI системите трябва да поддържат човешката автономия и вземането на решения.

Следователно те трябва да подкрепят

- Основни права
  - Оценка на правата на човека
- Човешка агенция.
  - Потребителите трябва да могат да вземат информирани автономни решения по отношение на AI системите.
- Човешки надзор.
  - Човешкият надзор помага да се гарантира, че дадена AI система не подкопава човешката автономност или причинява други неблагоприятни ефекти (подход „човек в цикъла“ (HITL), „човек в цикъла“ (HOTL) или „човек в командването“ (HIC) + публичен контрол)
- Техническа издръжливост и безопасност
  - Системите за AI да бъдат разработени с превантивен подход към рисковете и по такъв начин, че надеждно да се държат по предназначение, като същевременно минимизират непреднамерените и неочаквани вреди и предотвратяват неприемливи вреди.

# Човешка агенция и надзор (продължение)

- Устойчивост на атаки и сигурност
  - AI системите трябва да бъдат защитени срещу уязвимости, които могат да им позволят да бъдат използвани от противници
- Резервен план и обща безопасност
  - AI системите трябва да имат предпазни мерки, които позволяват резервен план в случай на проблеми
- Точност
  - Системите с AI трябва да имат способността да правят правилни преценки, например правилно да класифицират информацията в правилните категории, или способността да правят правилни прогнози, препоръки или решения въз основа на данни или модели.
- Надеждност и възпроизводимост.
  - Резултатите от AI системите трябва да бъдат възпроизводими, както и надеждни..



# Поверителност и управление на данните

- Предотвратяването на вреди налага поверителност и управление на данните:
  - Поверителност и защита на данните.
    - AI системите трябва да гарантират поверителност и защита на данните през целия жизнен цикъл на системата.
  - Качество и цялост на данните
    - Данните, използвани за обучение на системи, не трябва да съдържат социално конструирани пристрастия, неточности, грешки и не трябва да се добавят злонамерени данни.
  - Достъп до данни
    - Следва да бъдат въведени протоколи за данни, управляващи достъпа до данни.

# Прозрачност

- Това изискване е тясно свързано с принципа на обяснимостта
  - Проследимост.
    - Наборите от данни и процесите, които водят до решението на AI системата, трябва да бъдат документирани
  - Обяснимост.
    - Техническите процеси на AI система и свързаните с тях човешки решения трябва да бъдат обясними
  - Комуникация.
    - Хората имат право да бъдат информирани, че взаимодействат с AI система.

# Разнообразие, недискриминация и справедливост

- Трябва да позволим включването и разнообразието през целия жизнен цикъл на AI системата
- Избягване на несправедливи пристрастия
  - Предотвратяване на непреднамерени (не)преки предразсъдъци и дискриминация срещу определени групи или хора, което потенциално влошава предразсъдъците и маргинализацията, поради данни или алгоритми
- Достъпност и универсален дизайн.
  - AI системите трябва да бъдат ориентирани към потребителите и проектирани по начин, който позволява на всички хора да използват AI продукти или услуги, независимо от тяхната възраст, пол, способности или характеристики.
- Участие на заинтересованите страни.
  - Открита дискусия и участие на социални партньори и заинтересовани страни, включително широката общественост
- Разнообразие и приобщаващи дизайнерски екипи
  - екипите, които проектират, разработват, тестват и поддържат, внедряват и осигуряват тези системи, отразяват многообразието на потребителите и обществото като цяло

# Обществено и екологично благополучие

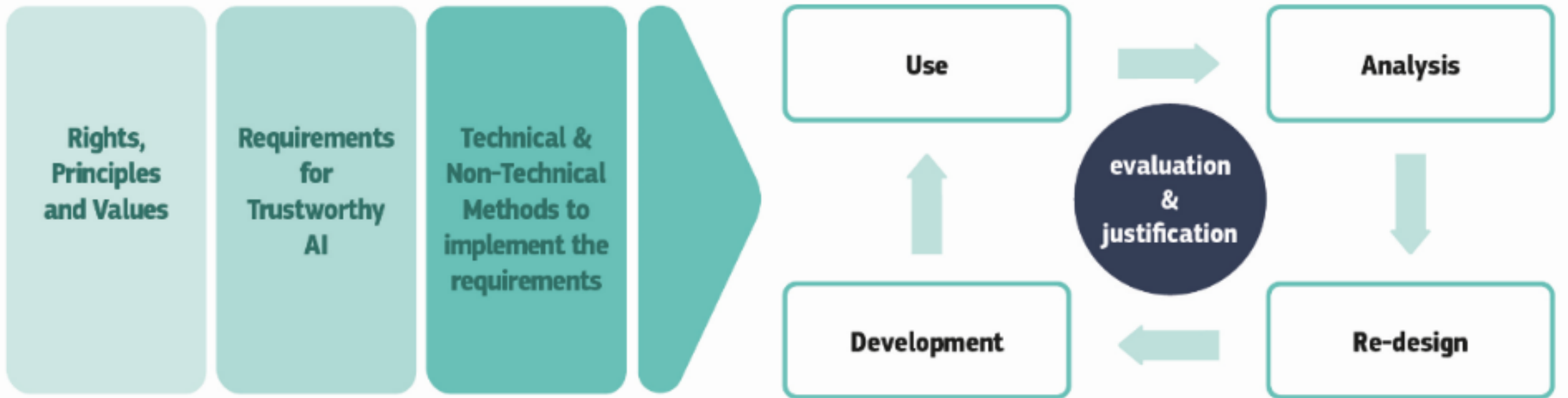
- По-широкото общество, други съзнателни същества и околната среда също трябва да се разглеждат като заинтересовани страни през целия жизнен цикъл на AI системата.
  - Устойчив и екологичен AI
    - Следва да се насърчават мерки, гарантиращи екологосъобразността на цялата верига на доставки на системите с AI.
  - Социално въздействие.
    - Поради това, въздействието на тези системи върху индивидите, групите и обществото трябва внимателно да се наблюдава и обмисля.
  - Общество и демокрация.
    - Да се взема предвид ефекта на AI върху институциите, демокрацията и обществото като цяло



# Отговорност

- Гарантиране на отговорност и отчетност за AI системите и техните резултати
  - Проверяемост
    - Възможност за оценка на алгоритми, данни и процеси на проектиране
  - Минимизиране и отчитане на негативните въздействия
    - Трябва да се осигури способността да се докладва за действия или решения, които допринасят за определен системен резултат, и да се реагира на последствията от такъв резултат.
  - Компромиси
    - Компромисите трябва да се разглеждат по рационален и методологичен начин в рамките на състоянието на техниката
  - Обезщетение.
    - Следва да се предвидят достъпни механизми, които да гарантират адекватно обезщетение

# Технически и нетехнически методи за реализиране на надежден AI



# Въпроси и предложения

- Въпроси
  - Документът Надежден AI предоставил ли ви е полезни указания?
  - Смятате ли, че са конкретно приложими?
  - Наистина ли са полезни етичните насоки, които не са правно обвързващи?
    - Някаква конкретна критика?
- Предложения
  - Прочетете целия документ!
  - Прочетете също

# Благодаря за вниманието

- [Giovanni.sartor@Unibo.it](mailto:Giovanni.sartor@Unibo.it)