MEANINGFUL HUMAN CONTROL

NATO C2COE seminar Get Connected

27 June 2019

Dr. Jurriaan van Diggelen

TNO innovation for life

Humans need to remain in control of AI; our AI systems must "do what we want them to do"

- January 2015

- Signed by >8000 AI experts
  - Stuart Russel,
  - Eric Horvitz
  - Elon Musk,
  - Stephen Hawking,
  - Nick Bostrom,
  - Geoffry Hinton,
  - …

An Open Letter

RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE

Click here to see this page in other languages: Chinese German Japanese Russian

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents – systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality – colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

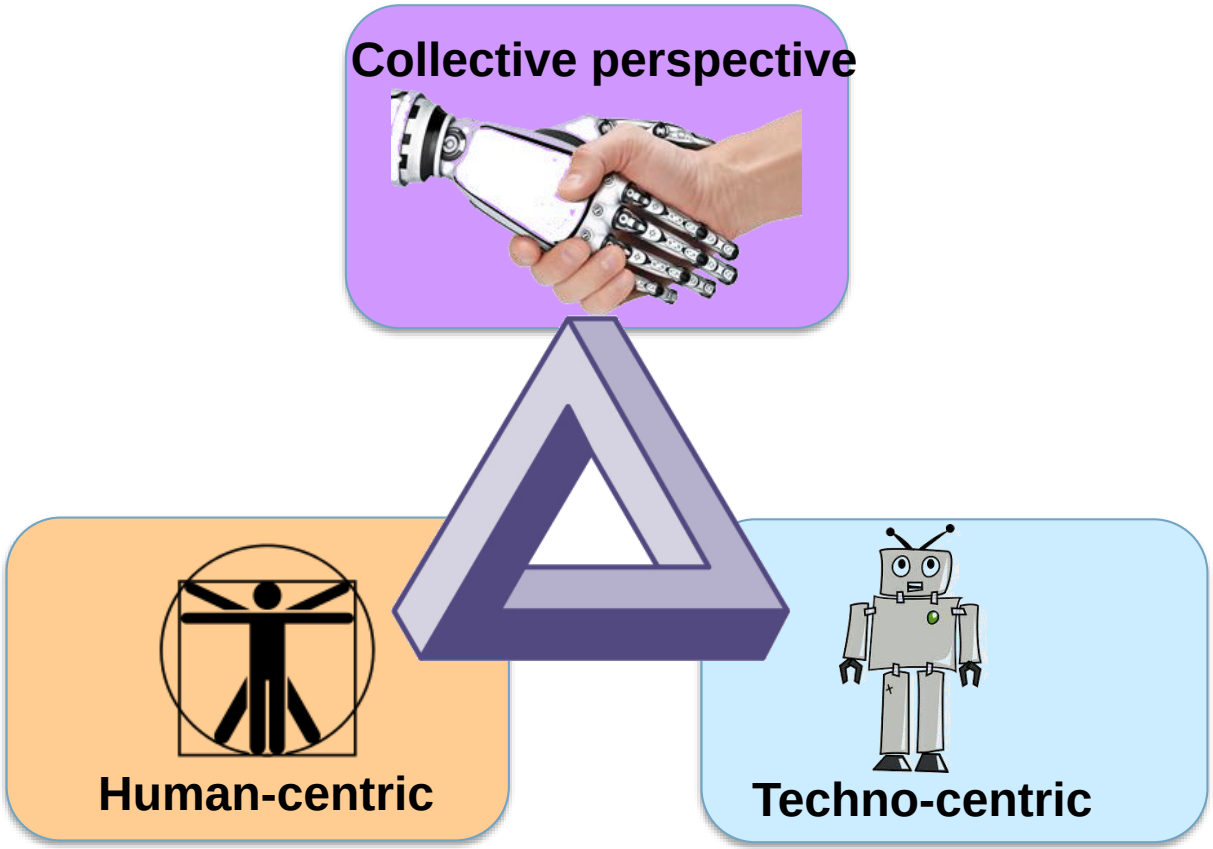
As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law and philosophy to computer security, formal methods and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

Part I
Three perspectives on AI

3 PERSPECTIVES ON ARTIFICIAL INTELLIGENCE



TECHNO-CENTRISM

I

SUFFICIENTLY DEVELOPED AI CAN BE APPLIED TO SOLVE ANY PROBLEM.

II

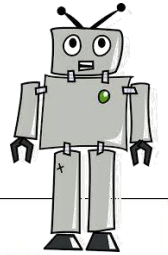
AI MIGHT INTRODUCE ADDITIONAL PROBLEMS, WHICH CAN IN TURN BE SOLVED BY AI.

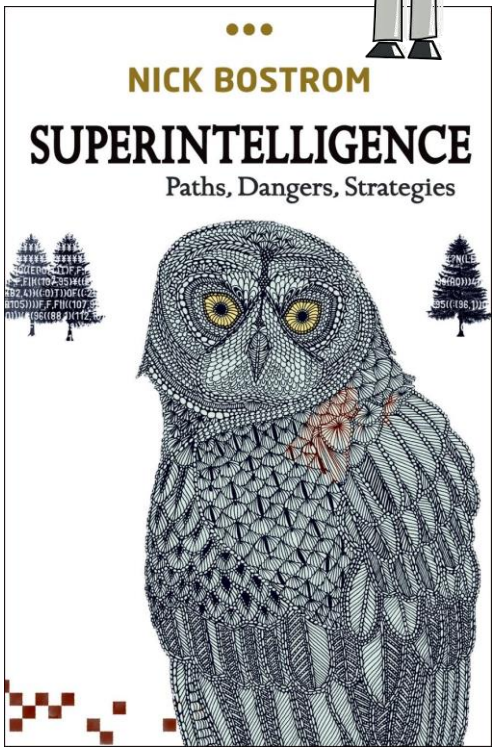
III

THE MORE AI IS DEVELOPED, THE LESS USER INTERACTION IS NEEDED.

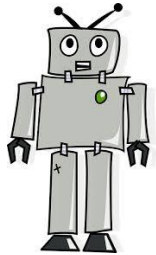
IV

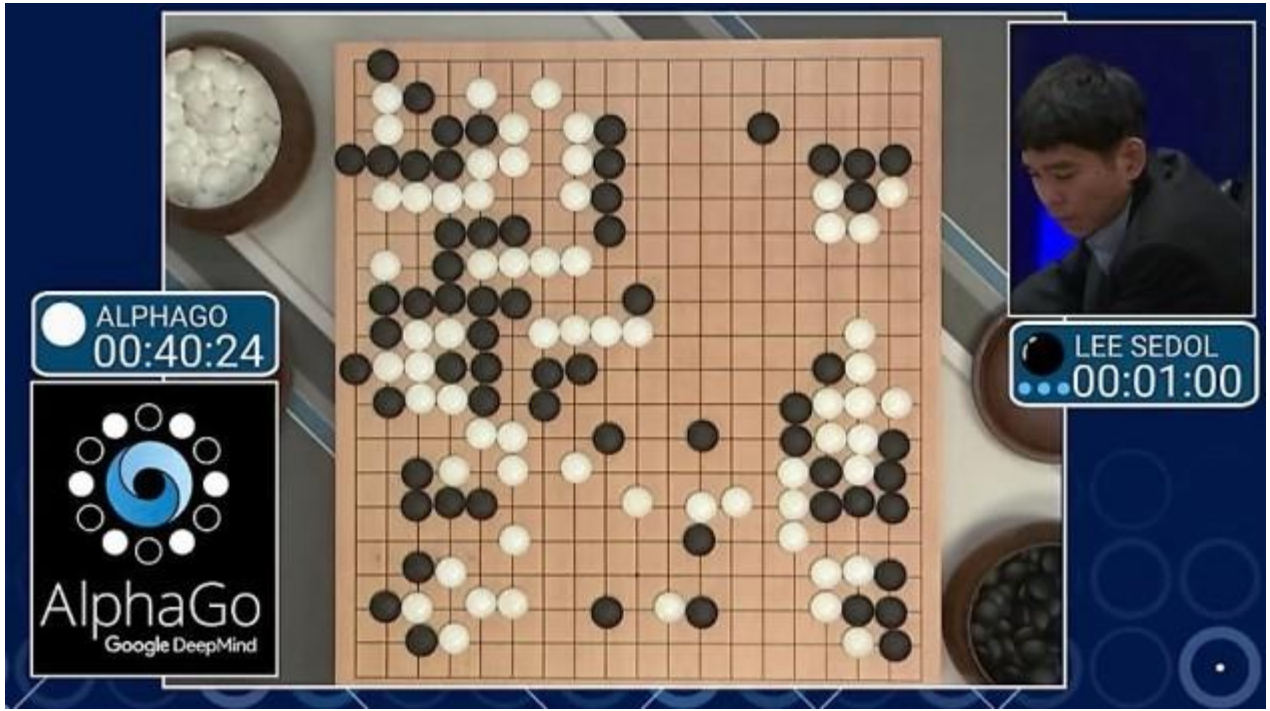
AI HAS VASTLY MORE POTENTIAL THAN HUMAN INTELLIGENCE



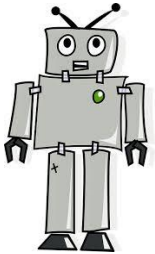


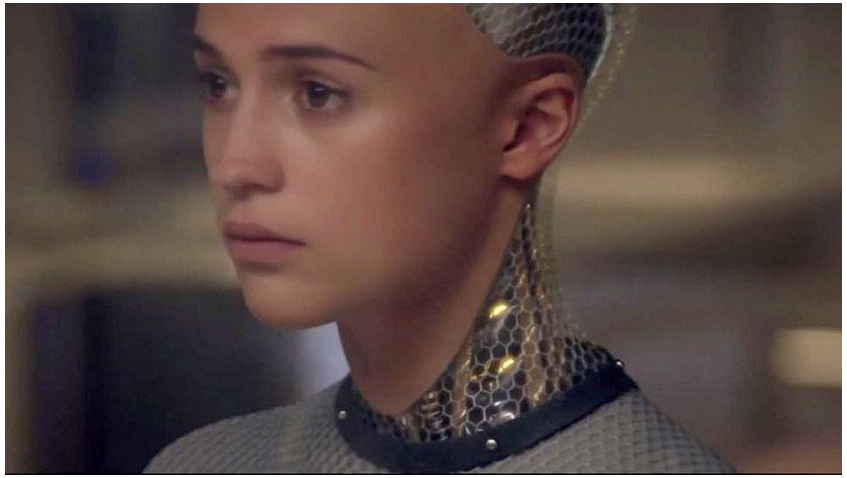
ALPHA (GO) ZERO





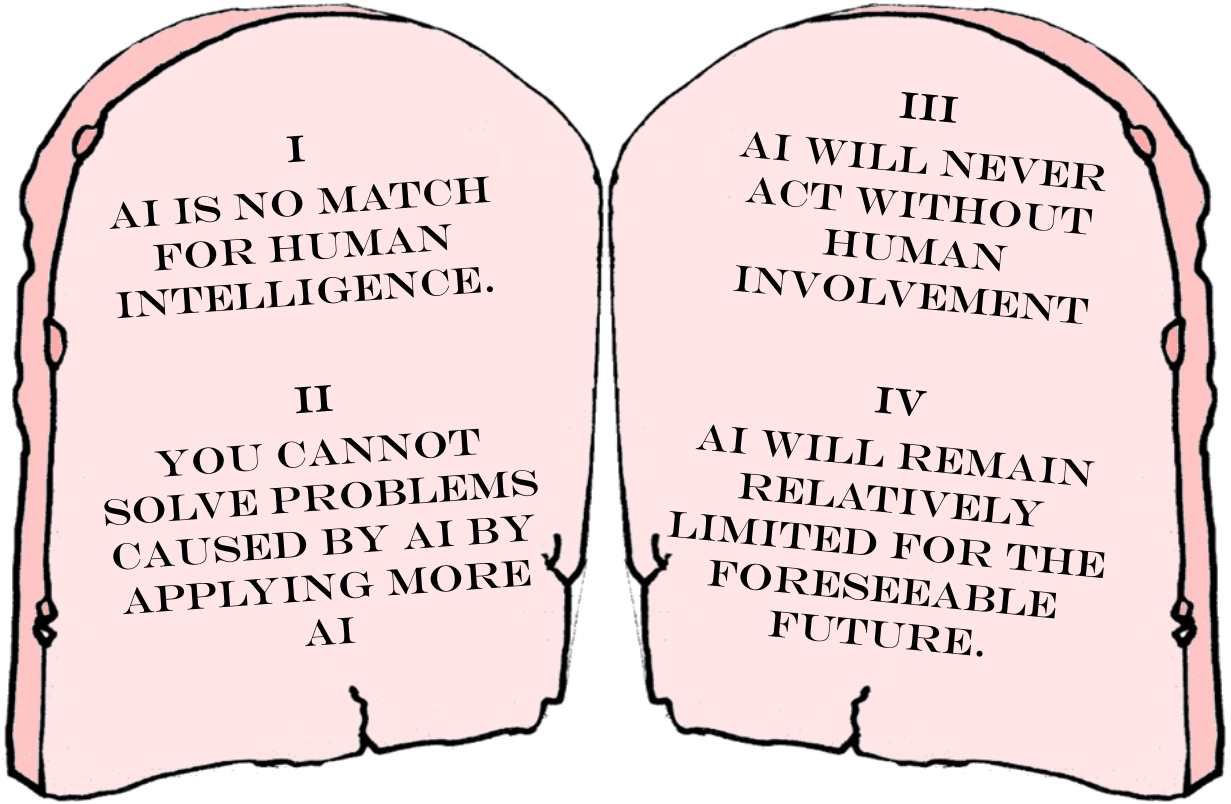
MEANINGFUL HUMAN CONTROL

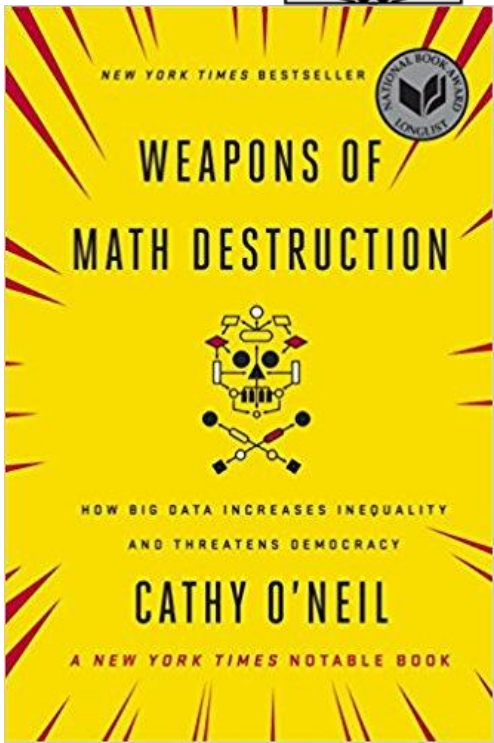




- **Problem:** How to control Artificial Super Intelligence.
- **Solution:** Program human values into AI system

HUMAN CENTRISM





PROPERTIES OF A WMD

- **Non-transparant**: It is unclear how AI arrives at its conclusions.
- **Scale**: The decisions made by AI affect large groups of people.
- **Damage:** The AI brings damage to large groups of people.

TEACHER ASSESSMENT TOOL

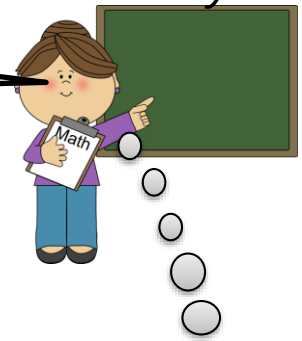
We must turn around underperforming schools in Washington D.C.

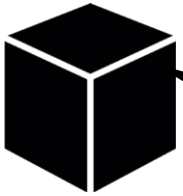
M50 MATHEMATICA Policy Research

We develop an objective and accurate model IMPACT to assess a teacher's performance

Sarah Wysocki

Along with 205 other teachers with a low IMPACT score, I got fired. Why?

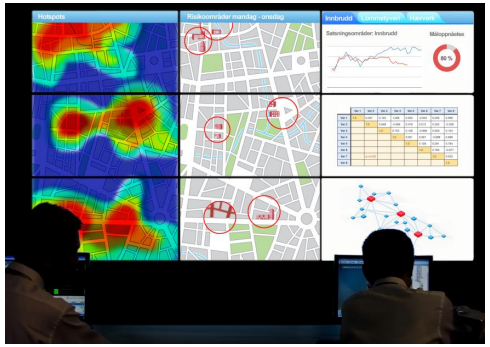




It's a complex algorithm you won't understand. Furthermore, it's corporate secret.

Many of my students came from a different school where they tampered test scores. They started scoring less in my tests…

OTHER EXAMPLES OF WMD'S



Predictive policing

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / 6 MONTHS AGO

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

Scan CV's

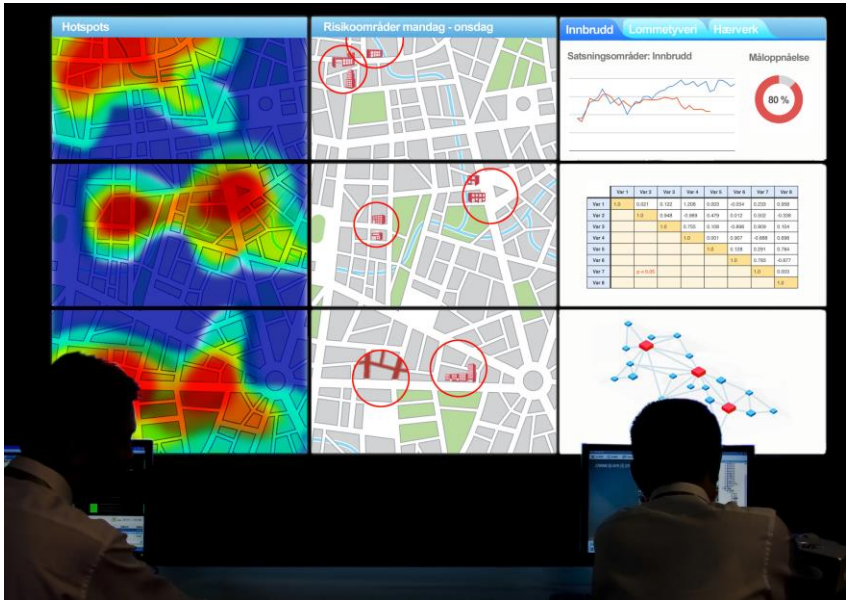
Political campaigning

Assess creditworthiness

Predict chance of recidivism

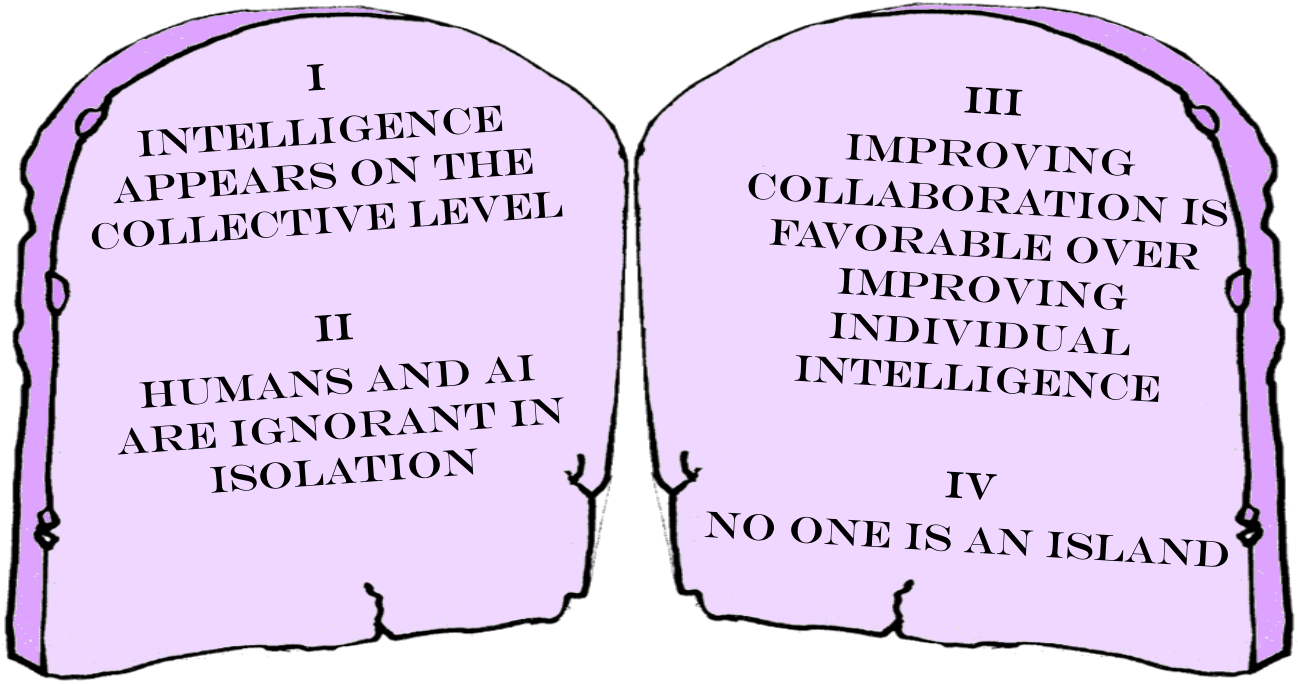
Calculate insurance premium

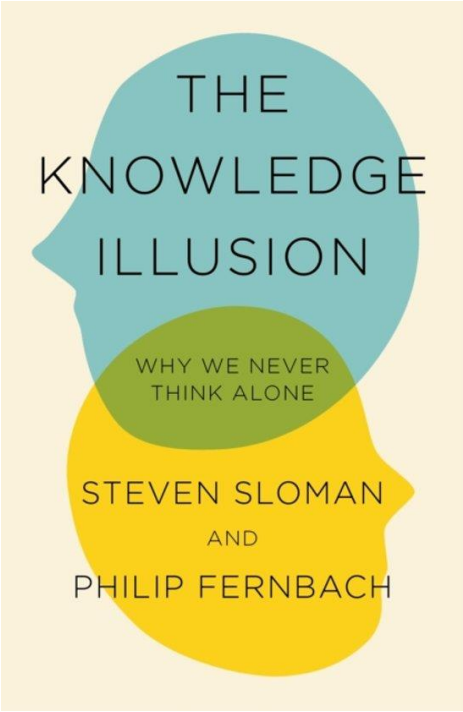
MEANINGFUL HUMAN CONTROL



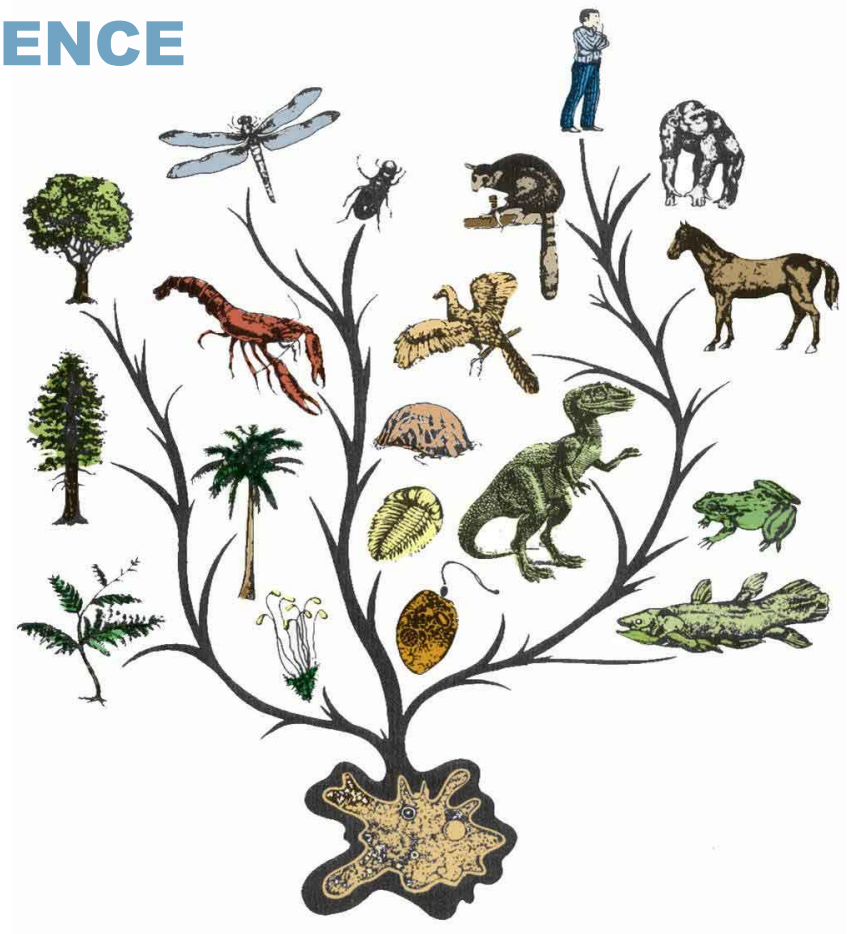
- **Problem:** Oversimplified AI models are granted too much control.
- **Solution:** Apply AI sparingly.

COLLECTIVE INTELLIGENCE





SOCIAL INTELLIGENCE



SOCIAL AI IS ESSENTIAL


GOVERNMENT
REGULATIONS
Pizza
deliveroo
AI Scheduler

MEANINGFUL HUMAN CONTROL

Problem: Integrating AI into teams, organisations, and society inevitably disturbs the equilibrium between autonomy and control.

Solution: Detect and redirect undesirable developments.

Part II
AI in defense

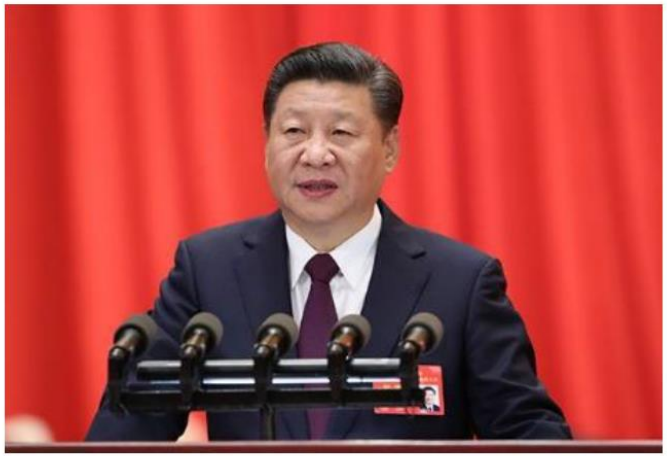
US & WORLD | TECH | ARTIFICIAL INTELLIGENCE

Putin says the nation that leads in AI 'will be the ruler of the world'

43

The Russian president warned that artificial intelligence offers 'colossal opportunities' as well as dangers

By James Vincent | @jjvincent | Sep 4, 2017, 4:53am EDT



"… by 2030, China's AI theories, technologies, and applications should achieve world leading levels, making China the world's primary AI innovation center…"

6 VOORALSNOG LIJKT CHINA DE RACE TE WINNEN

Deze mondiale race om de beste en meeste AI speelt zich vooral af tussen de VS en China, waar China *en route* is om gaan winnen. Dit blijkt uit een vergelijking van Nederland en/of Europa met China en de VS op drie vlakken.

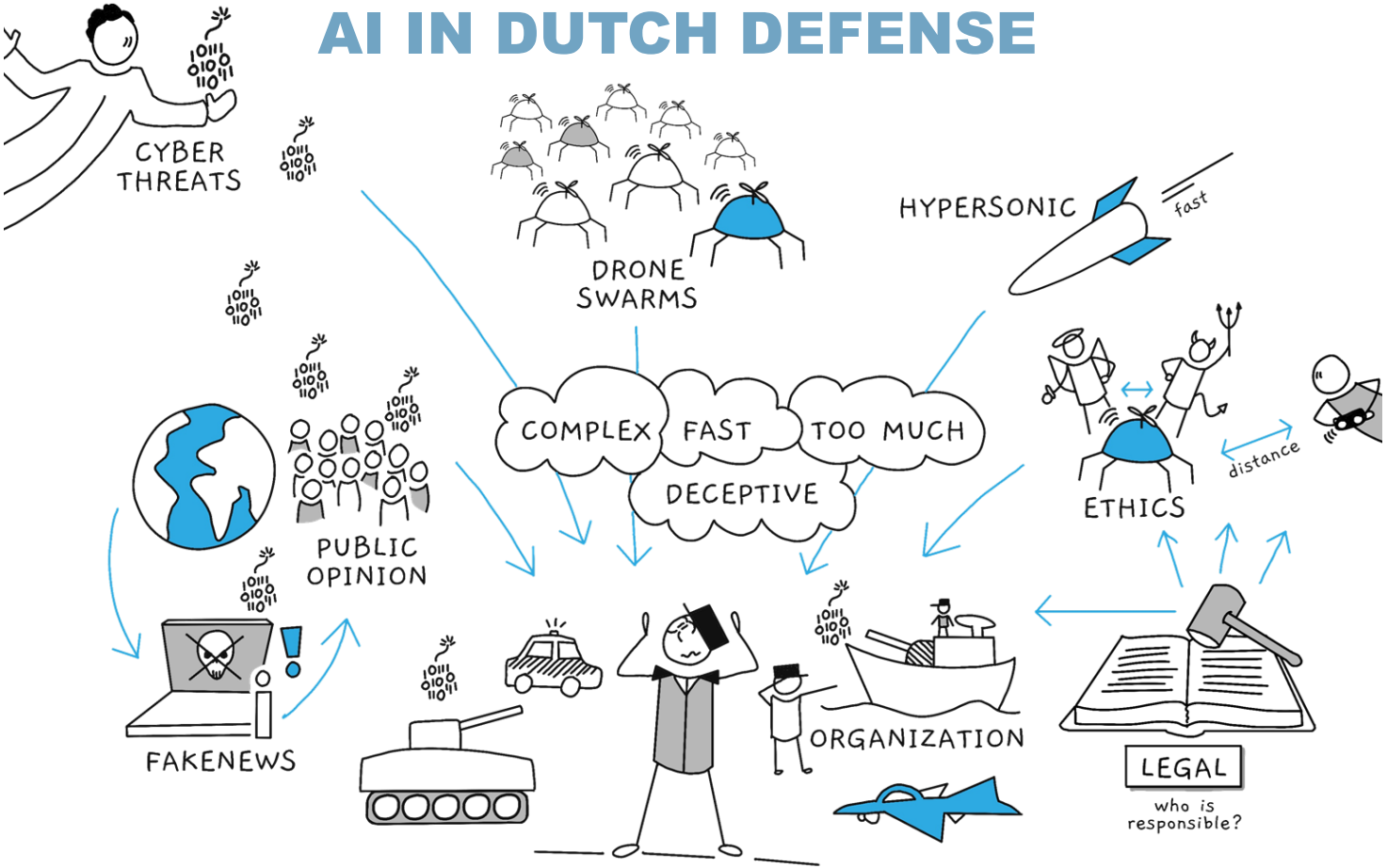
Het onderwijs in China is weliswaar kwalitatief wat minder goed dan dat van de VS en Europa, maar China compenseert dat op twee manieren. Ten eerste heeft China een relatief hoog aantal studenten in dit vakgebied. Op haar beste drie universi-

Translated from Chinese AI Development Plan, July 2017

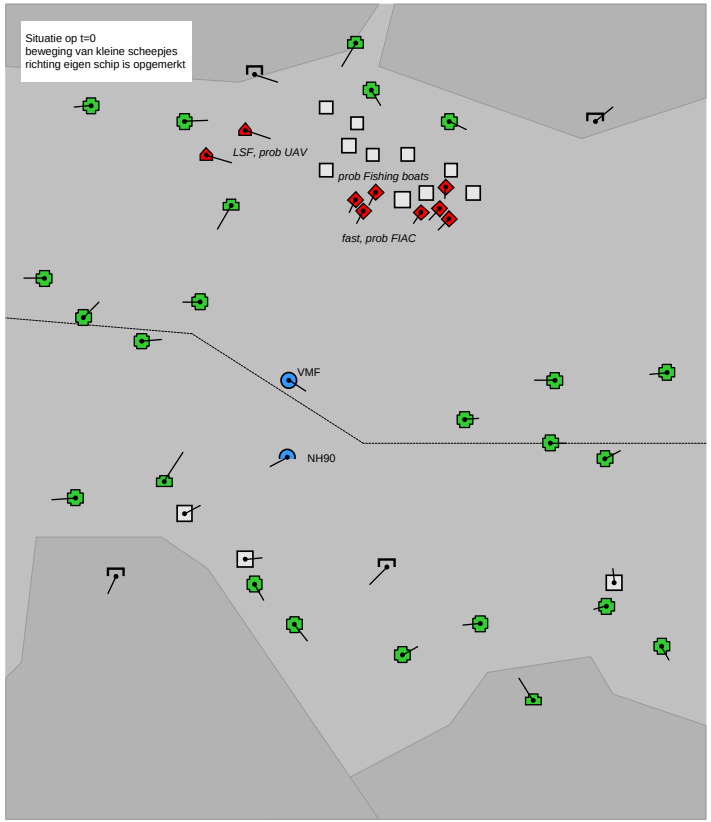
"And we believe quite strongly that the technological sauce of the Third Offset is going to be advances in Artificial Intelligence (AI) and autonomy."

Former U.S. Deputy Secretary of Defense Bob Work, April 2016

AI IN DUTCH DEFENSE



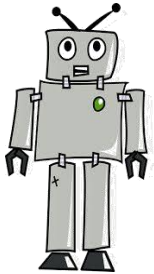
SCENARIO: SWARM ATTACK



Unpredictable
Manipulative
Time critical
Ethical
Information overload

WHAT IS OUT THERE NOW?

MEANINGFUL HUMAN CONTROL OVER LAWS



The battle is too fast to allow human involvement, so we need autonomous AI!
Artificial ethics can make war more humane.

MEANINGFUL HUMAN CONTROL OVER LAWS

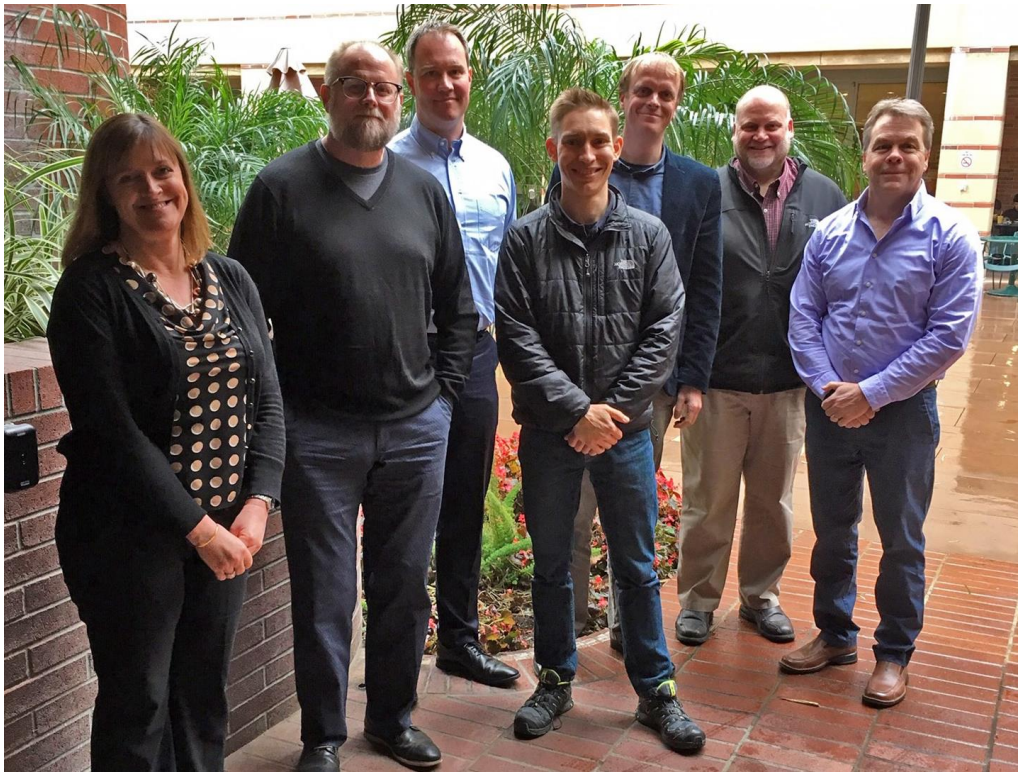
Prohibit all LAWS!

Article 36

Humans should exercise control over individual attacks, not simply overall operations. Only by prohibiting the use of fully autonomous weapons can such control be guaranteed.

NATO HFM-ET-178: MHC OVER AI BASED SYSTEMS

AFRL
TNO
DSTL
NASA
Fraunhofer
FOI



WHAT IS MEANINGFUL HUMAN CONTROL?

TNO innovation for life

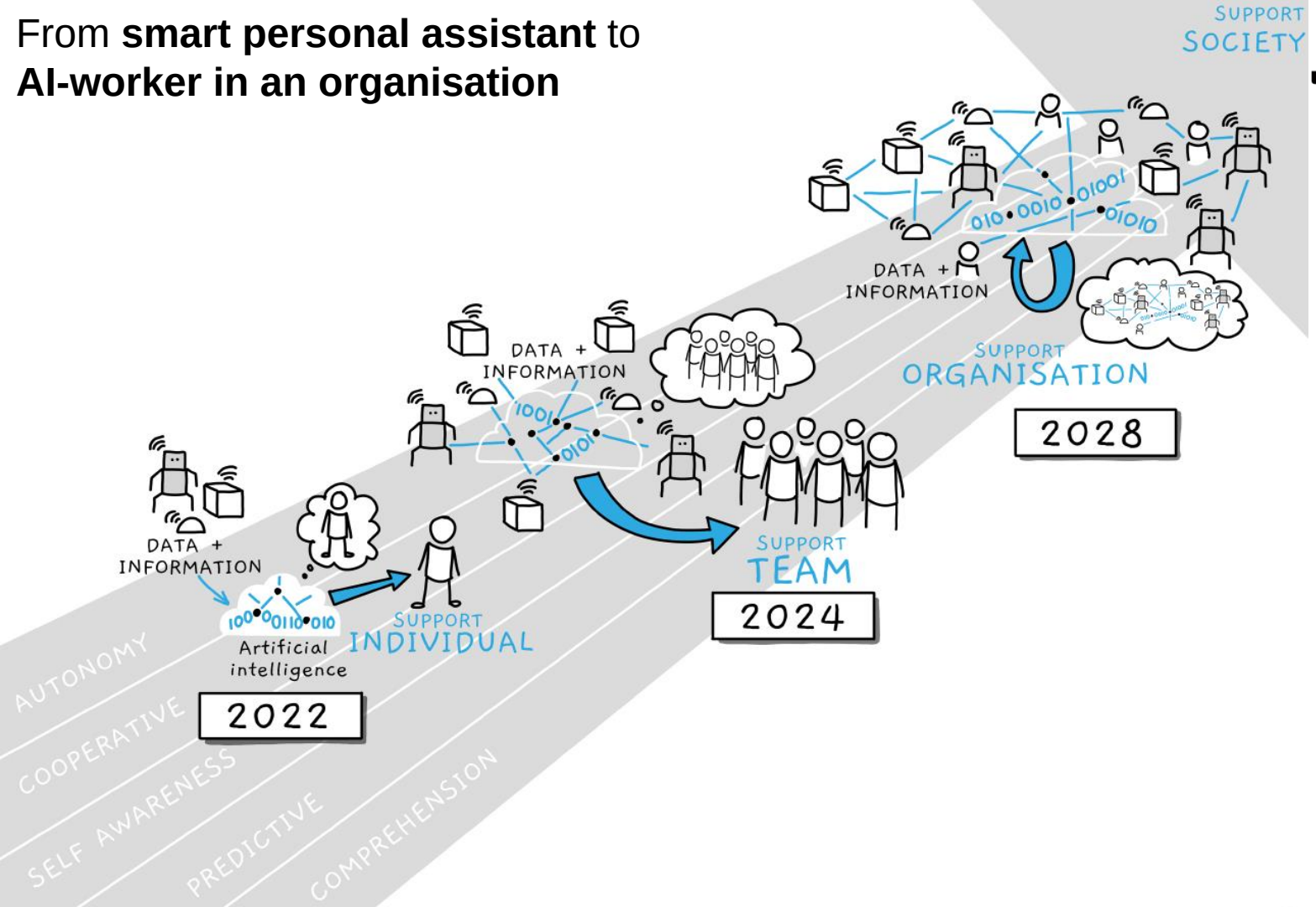
Humans have the ability to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment.

Characteristics:

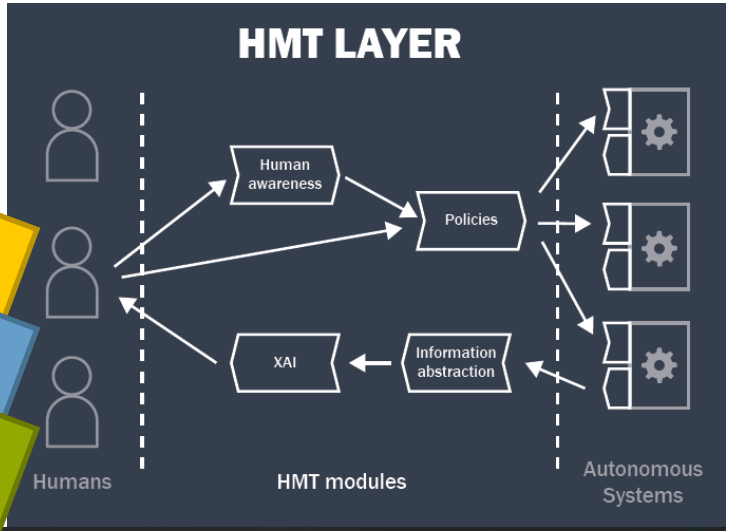
- Human has freedom of choice
- Human has ability to impact the behaviour of the system
- Human has time to decide to engage and sufficient situation, and system understanding
- Human is capable to predict the behavior of the system and the effects of the environment (physical and information)
- Influence over AI-based systems can be achieved in various ways, such as policy-making, training, HMI design, organizational design, operations, etc.
- The above encompasses cases from instantaneous (e.g. number of seconds) to very delayed response (several hours to days, e.g. before-the-loop) to control.

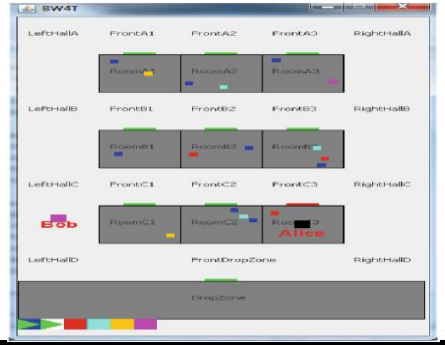
Part III
Human machine teaming

From **smart personal assistant** to **AI-worker in an organisation**



HMT AT TNO

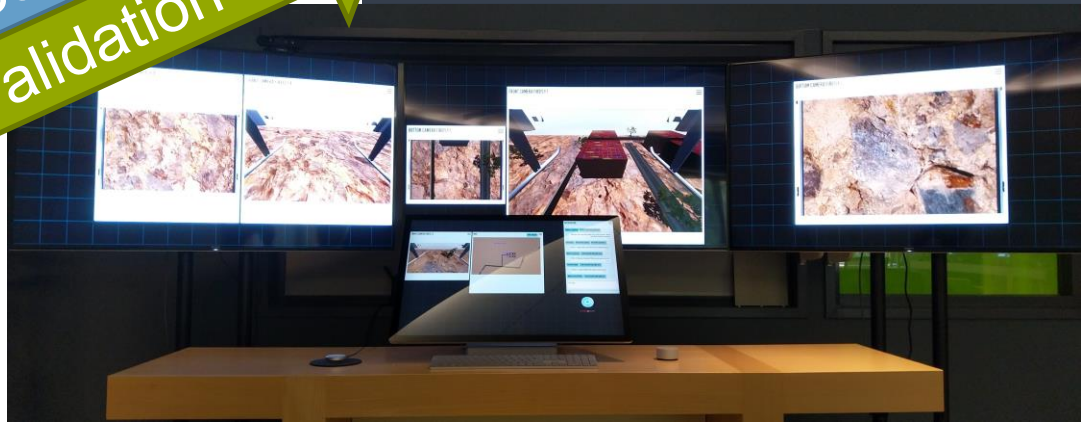




HMT specification

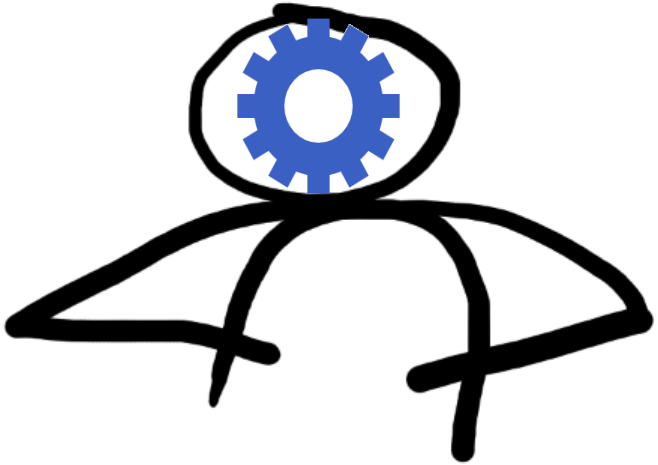
HMT technology

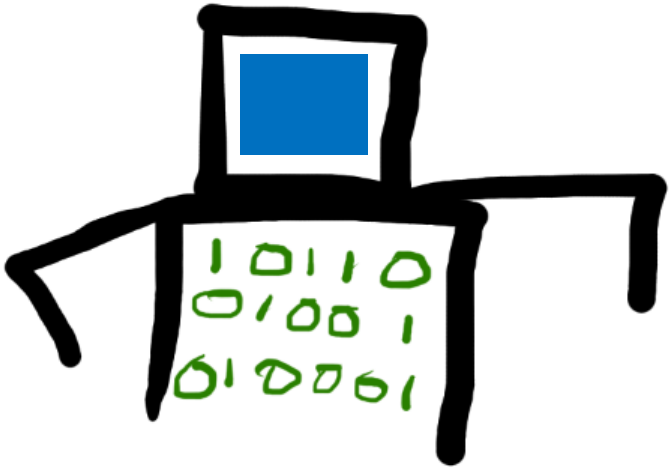
HMT validation



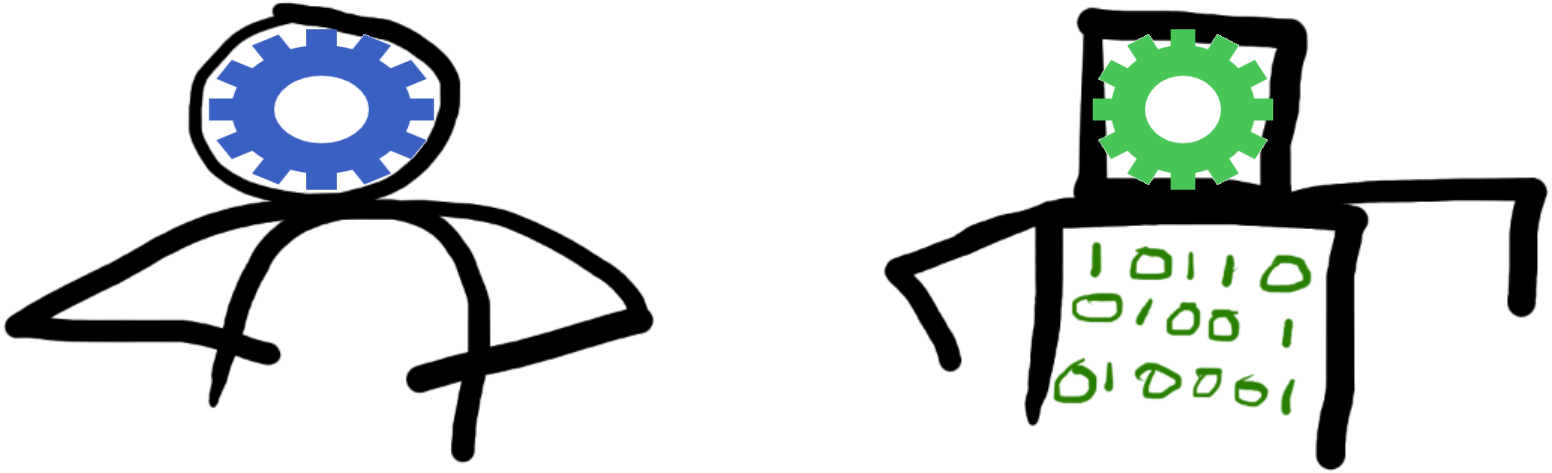
XAI

COMPUTER AS A TOOL

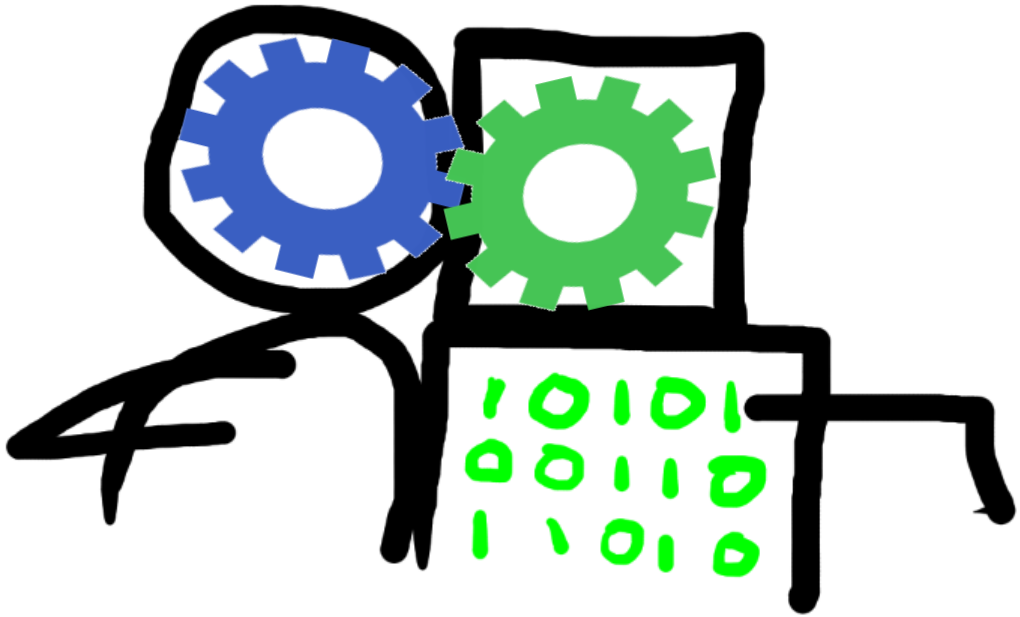




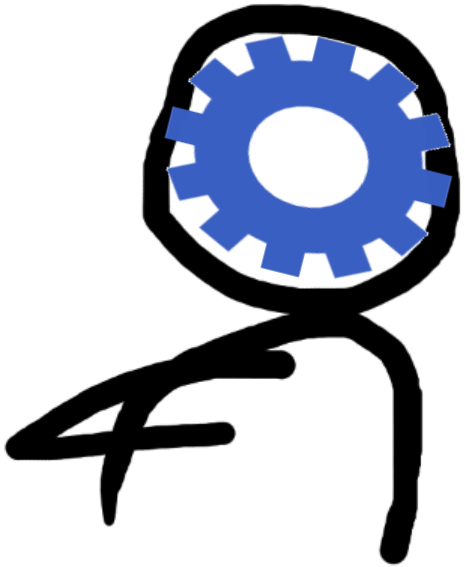
COMPUTER AS AN ISOLATED AGENT

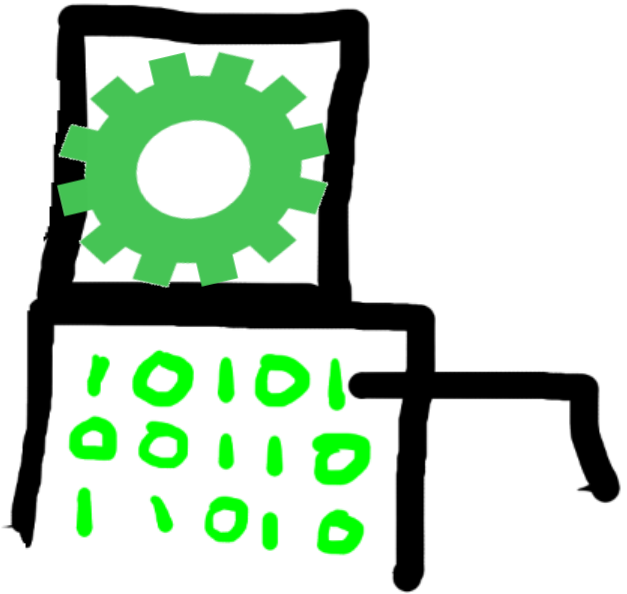


COMPUTER AS A TEAMMATE

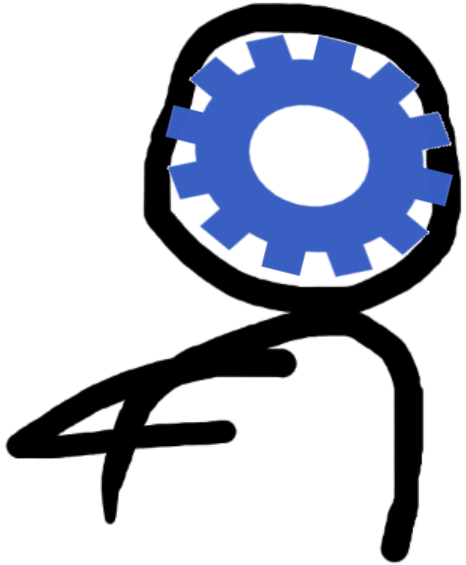


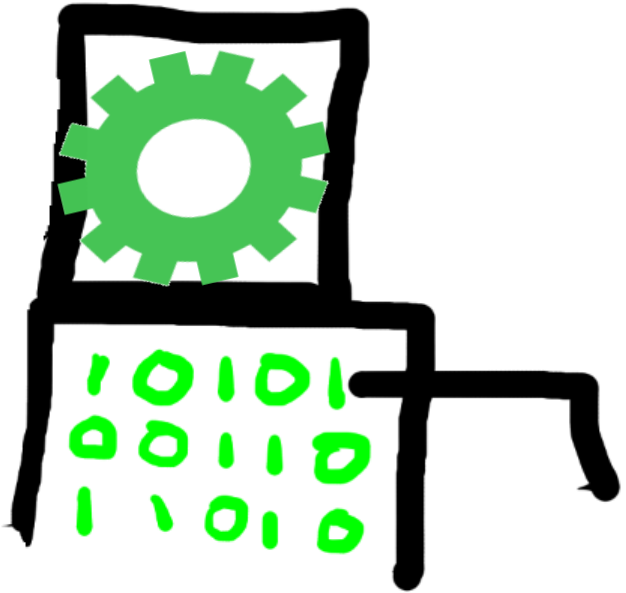
DYNAMIC TEAMING



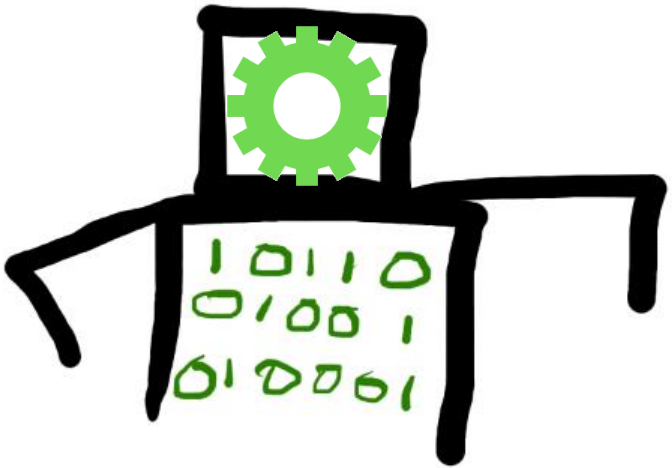


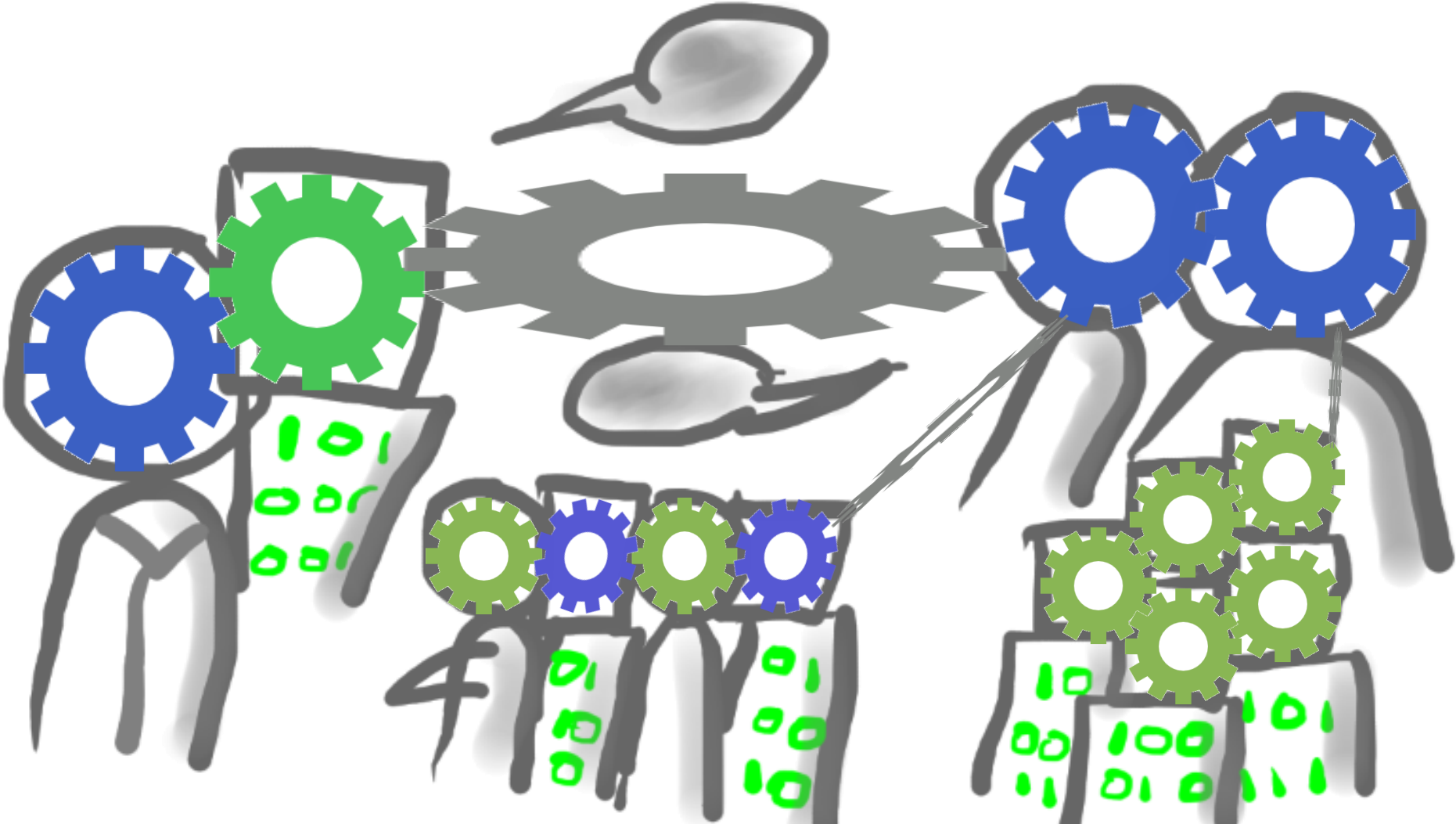
DYNAMIC TEAMING

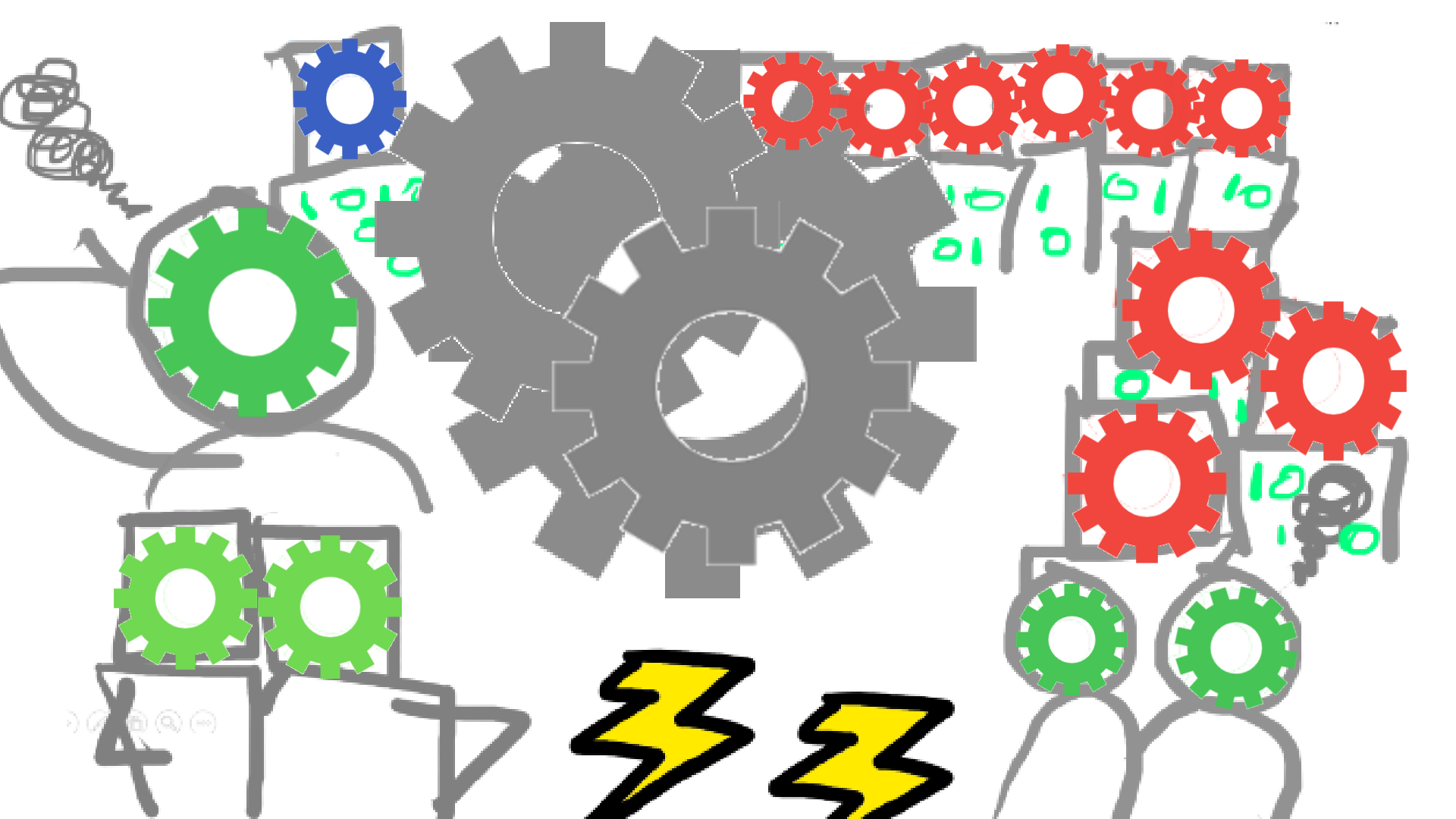




RUNAWAY AI



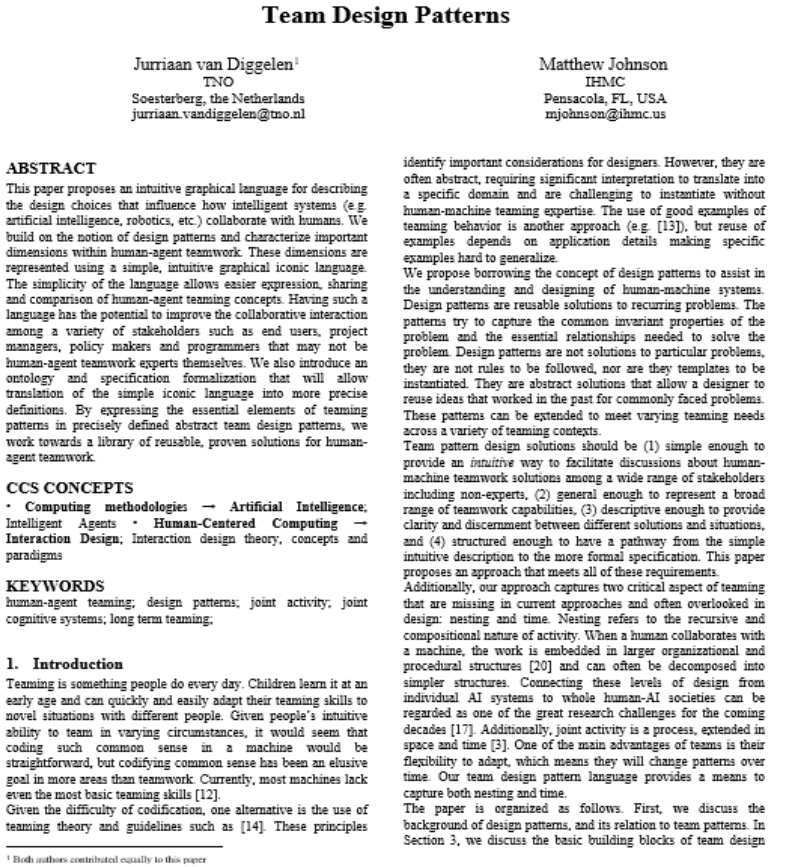




TEAM DESIGN PATTERNS

- How to design coherent human agent teams in a way that is
  - **Simple and intuitive** to allow communication among stakeholders
  - **General** enough to represent a broad range of teamwork
  - **Descriptive** enough to allow comparison of different solutions and situations
  - **Structured** enough to have a pathway from the simple intuitive description to the more formal specification.

- Fosus on:
  - **Nesting**
  - **Time**



Team Design Patterns

Jurriaan van Diggelen[1]
TNO
Soesterberg, the Netherlands
jurriaan.vandiggelen@tno.nl

Matthew Johnson
IHMC
Pensacola, FL, USA
mjohnson@ihmc.us

ABSTRACT

This paper proposes an intuitive graphical language for describing the design choices that influence how intelligent systems (e.g. artificial intelligence, robotics, etc.) collaborate with humans. We build on the notion of design patterns and characterize important dimensions within human-agent teamwork. These dimensions are represented using a simple, intuitive graphical iconic language. The simplicity of the language allows easier expression, sharing and comparison of human-agent teaming concepts. Having such a language has the potential to improve the collaborative interaction among a variety of stakeholders such as end users, project managers, policy makers and programmers that may not be human-agent teamwork experts themselves. We also introduce an ontology and specification formalization that will allow translation of the simple iconic language into more precise definitions. By expressing the essential elements of teaming patterns in precisely defined abstract team design patterns, we work towards a library of reusable, proven solutions for human-agent teamwork.

CCS CONCEPTS

• **Computing methodologies** → **Artificial Intelligence**; Intelligent Agents • **Human-Centered Computing** → **Interaction Design**; Interaction design theory, concepts and paradigms

KEYWORDS

human-agent teaming; design patterns; joint activity; joint cognitive systems; long term teaming;

1. Introduction

Teaming is something people do every day. Children learn it at an early age and can quickly and easily adapt their teaming skills to novel situations with different people. Given people's intuitive ability to team in varying circumstances, it would seem that coding such common sense in a machine would be straightforward, but codifying common sense has been an elusive goal in more areas than teamwork. Currently, most machines lack even the most basic teaming skills [12].

Given the difficulty of codification, one alternative is the use of teaming theory and guidelines such as [14]. These principles identify important considerations for designers. However, they are often abstract, requiring significant interpretation to translate into a specific domain and are challenging to instantiate without human-machine teaming expertise. The use of good examples of teaming behavior is another approach (e.g. [13]), but reuse of examples depends on application details making specific examples hard to generalize.

We propose borrowing the concept of design patterns to assist in the understanding and designing of human-machine systems. Design patterns are reusable solutions to recurring problems. The patterns try to capture the common invariant properties of the problem and the essential relationships needed to solve the problem. Design patterns are not solutions to particular problems, they are not rules to be followed, nor are they templates to be instantiated. They are abstract solutions that allow a designer to reuse ideas that worked in the past for commonly faced problems. These patterns can be extended to meet varying teaming needs across a variety of teaming contexts.

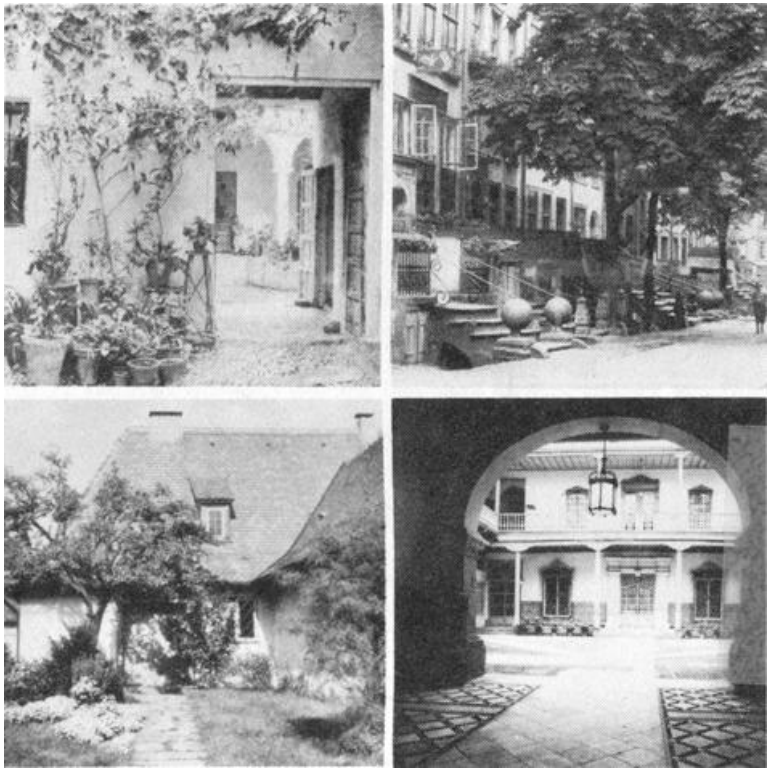
Team pattern design solutions should be (1) simple enough to provide an *intuitive* way to facilitate discussions about human-machine teamwork solutions among a wide range of stakeholders including non-experts, (2) general enough to represent a broad range of teamwork capabilities, (3) descriptive enough to provide clarity and discernment between different solutions and situations, and (4) structured enough to have a pathway from the simple intuitive description to the more formal specification. This paper proposes an approach that meets all of these requirements.

Additionally, our approach captures two critical aspect of teaming that are missing in current approaches and often overlooked in design: nesting and time. Nesting refers to the recursive and compositional nature of activity. When a human collaborates with a machine, the work is embedded in larger organizational and procedural structures [20] and can often be decomposed into simpler structures. Connecting these levels of design from individual AI systems to whole human-AI societies can be regarded as one of the great research challenges for the coming decades [17]. Additionally, joint activity is a process, extended in space and time [3]. One of the main advantages of teams is their flexibility to adapt, which means they will change patterns over time. Our team design pattern language provides a means to capture both nesting and time.

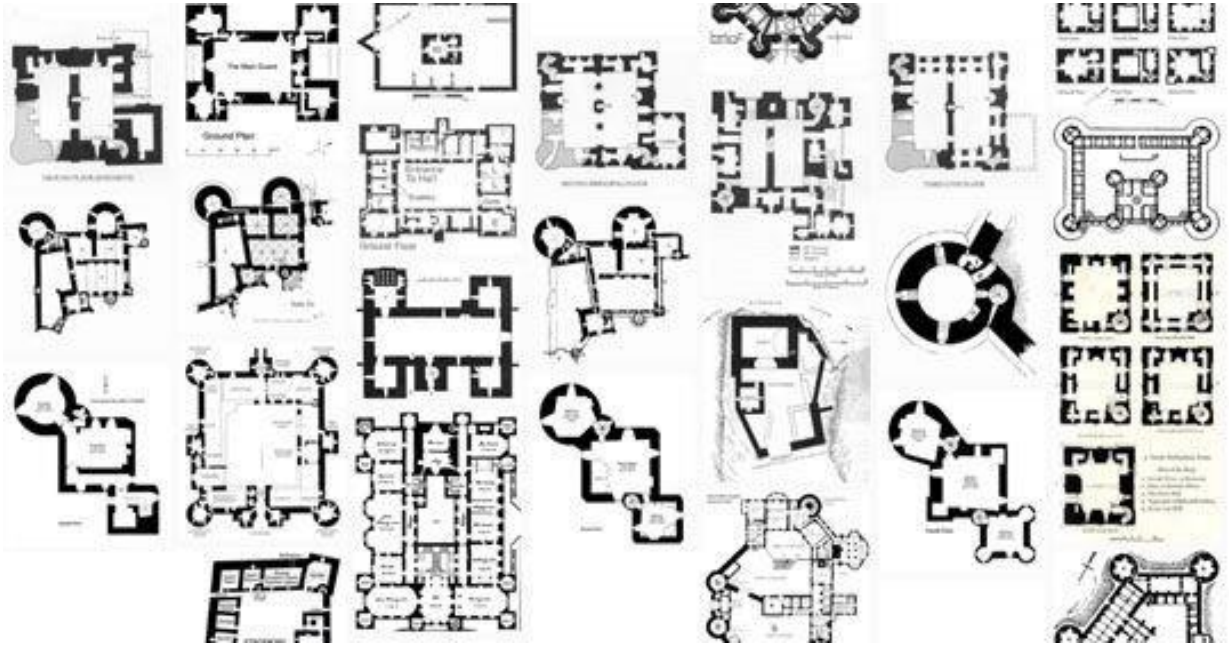
The paper is organized as follows. First, we discuss the background of design patterns, and its relation to team patterns. In Section 3, we discuss the basic building blocks of team design

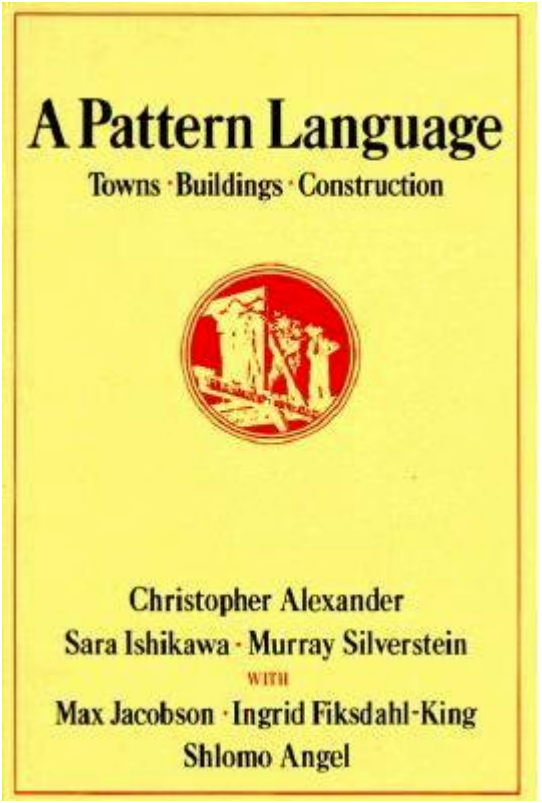
[1] Both authors contributed equally to this paper

CHRISTOPHER ALEXANDER

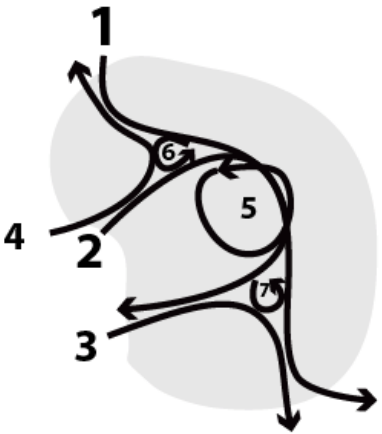


A PATTERN LANGUAGE





Christopher Alexander's Default Design Approach

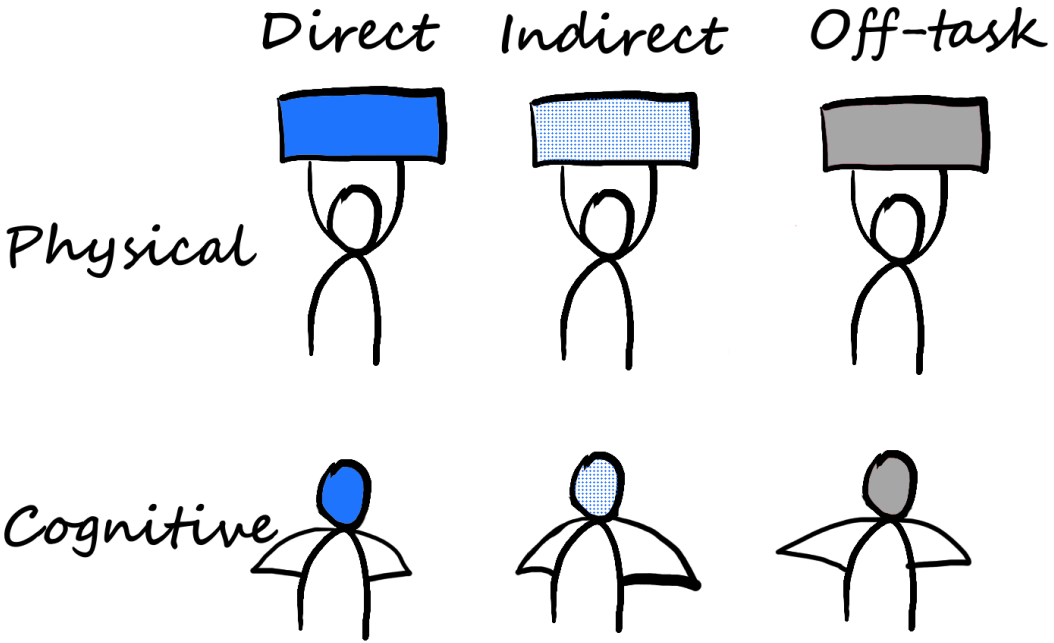


a. Start with a whole...

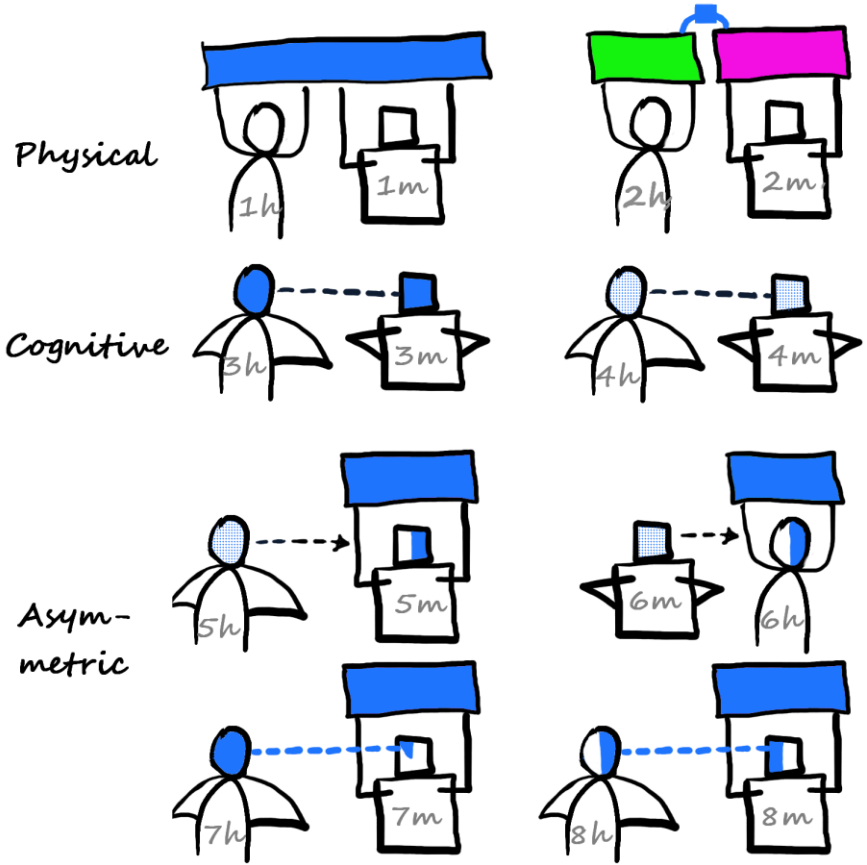
b. Differentiate it...

c. Into parts...

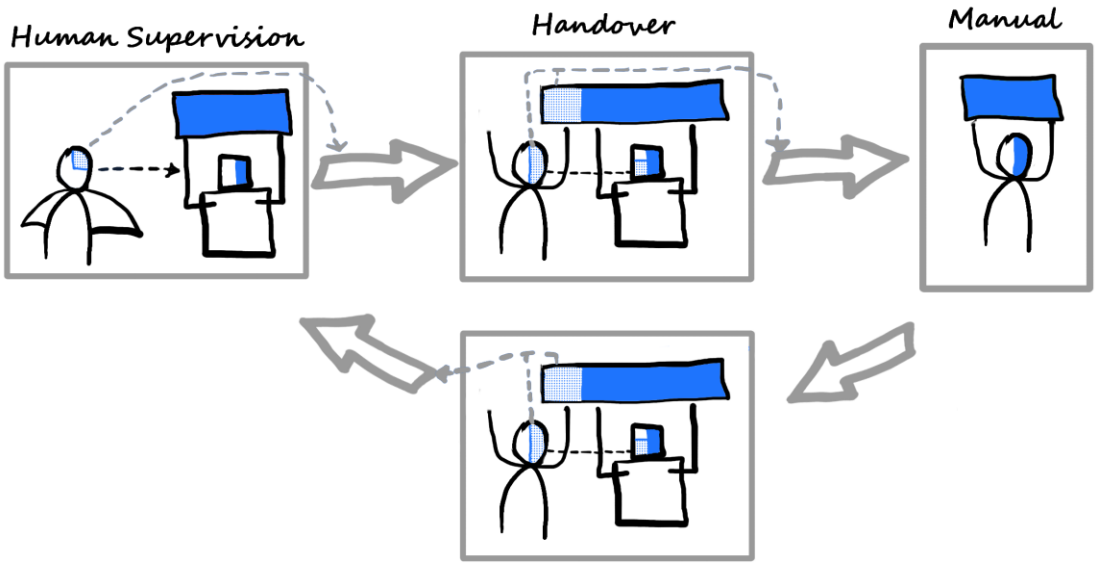
BASIC TYPES OF WORK



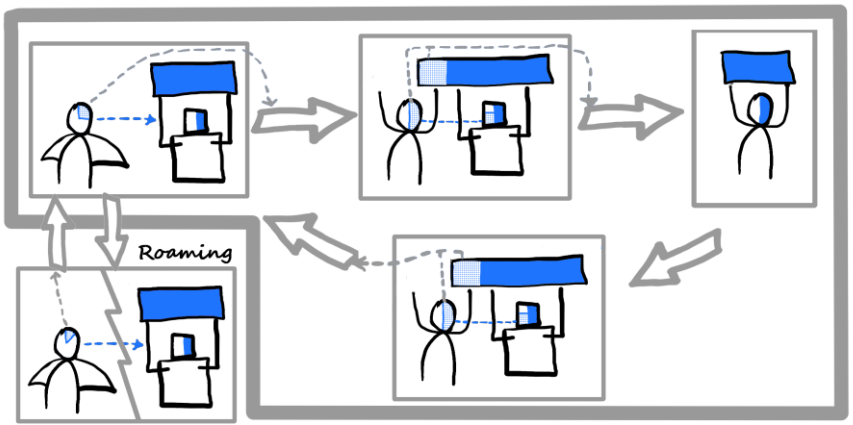
JOINT WORK

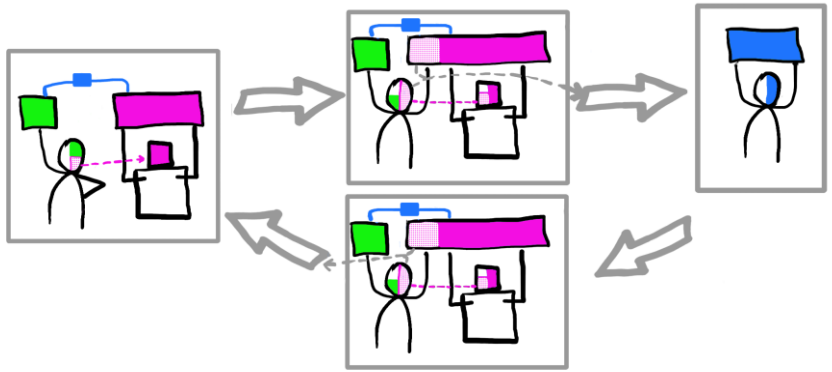

Physical
1h
1m
2h
2m
Cognitive
3h
3m
4h
4m
Asym-
metric
5h
5m
6m
6h
7h
7m
8h
8m

SUPERVISORY CONTROL

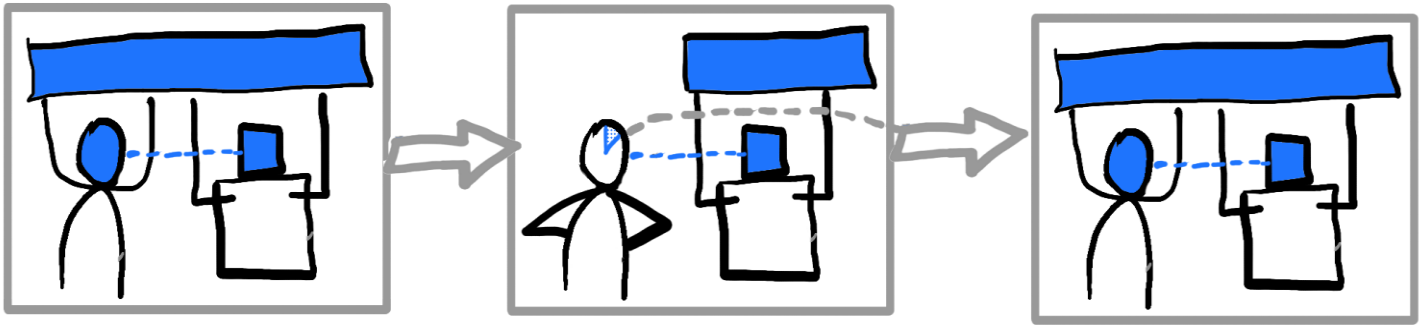


VARIANTS OF SUPERVISORY CONTROL





HIGHLY AUTONOMOUS PATTERNS



FORMAL SPECIFICATION

Teleoperation : **Team Design Pattern**

Name : "Tele operation"
Image : Img1
Use when : "machine has limited autonomous capability, and human skilled operators are available..."
Positive effect: "Clear single point of control at human operator"
Negative effect : "Imposes heavy taskload on the human"
Example : "Teleoperation of a UAV..."
Involves actors : [7h,7m]

7h : **Human**

Name: "7h"
Performs <level of engagement = 1.0> Teleoperating

Teleoperating: **Direct Cognitive Work**

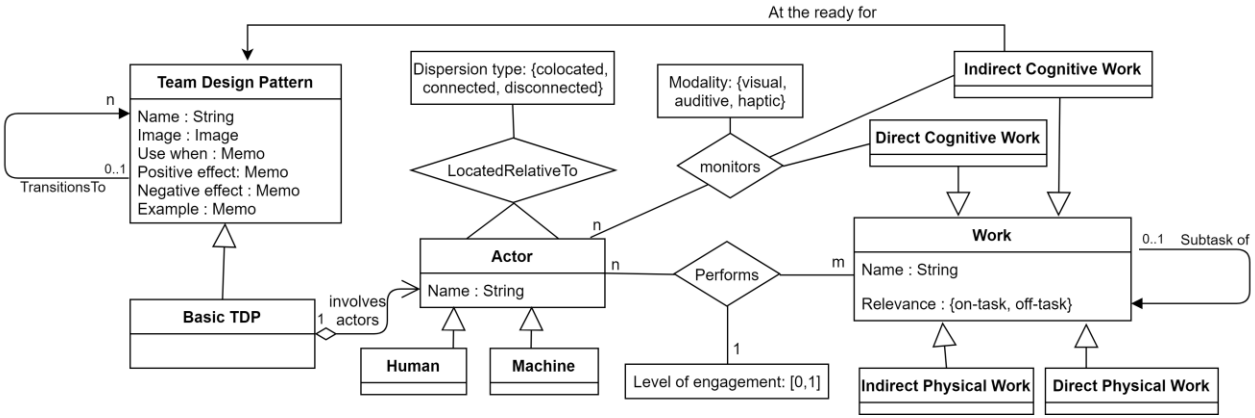
Name: "teleoperating"
Relevance = "on-task
Monitors <modality = auditive> [7h,7m]

7m : **Machine**

Name: "7m"
Performs <level of engagement = 0.1> Teleoperating
Performs <level of engagement = 1.0> PerformInstructions

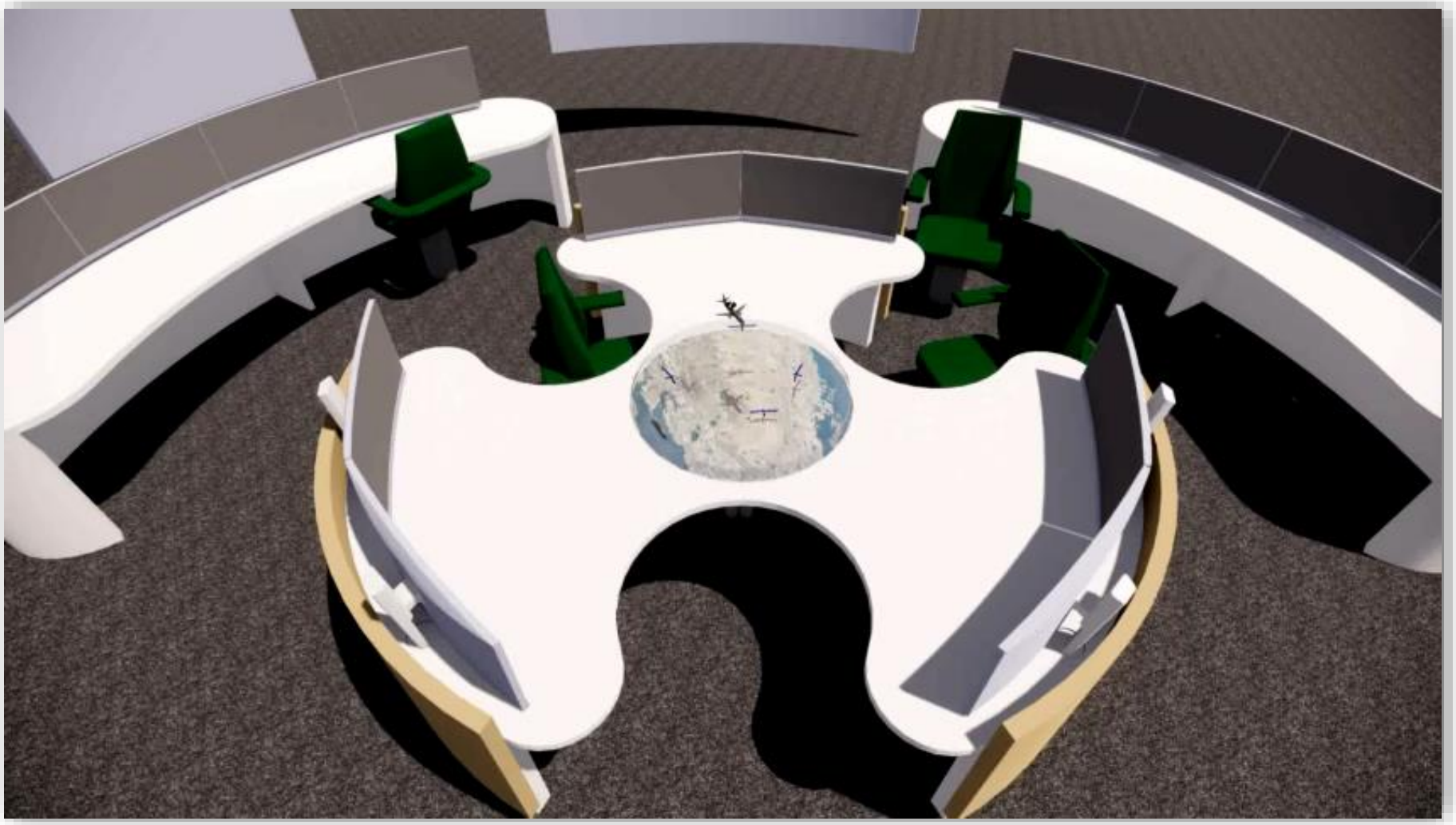
PerformInstructions: **Direct physical Work**

Name: "PerformInstructions"
Relevance = "on-task



GOAL: develop a pattern library for meaningful human control.

SUPPORT MULTI TEAM SYSTEMS



"The winner of the robotics revolution will not be who develops this technology first or even who has the best technology, but who figures out how to best use it."

Paul Scharre,
Robotics on the Battlefield Part 1: Range, Persistence and Daring


ARTIFICIAL INTELLIGENCE
NATURAL INTELLIGENCE
QUESTIONS?
SUMANTA BARUAH