AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint

AI-MAN (ICT-38) Projects Cluster
Workshops Series (On-Line)
Monday, November 25th 2021

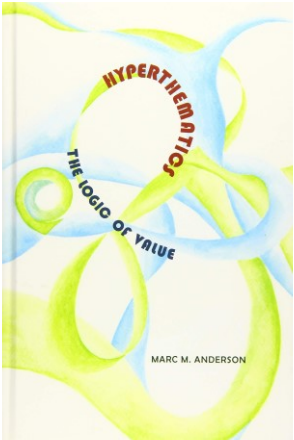
“Ethical and Legal Issues of Artificial Intelligence In Manufacturing”

This product is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 957391.

AI-Proficient Ethics Team

- Dr Karën Fort (Ethics Team leader)
    - Maître de Conference, LORIA, UMR 7503, Université de Lorraine, Inria and CNRS
    - AI Ethics Specialist
    - Member of multiple AI Ethics Committees

- Dr Marc Anderson (post-doc hired by the project for 1 year + 1 year)
    - Philosopher
    - Research in AI Ethics + Ethics as Value, applied to products and manufacturing
        - *Hyperthematics: The Logic of Value* (2019)



Human in the Loop in AI-Proficient

- Human-in-the-loop, Human-on-the-loop, and Human-in-command were terms integrated in AI-Proficient from the beginning
- Use Cases proposed by manufacturing partners requested human involvement with AI based on these terms (for one partner 10 UCs proposed: 8 HIC, 1 HITL, 1 HIC/or HITL)
- Often ambiguity in use of terms

  Examples: *HIC human is specified but the degree of command is unclear*;

  no-one is specified in the UC as the HIC

HITL Terms: A Deeper Problem

- EU Commission Ethics Guidelines for Trustworthy AI (HLEG on AI, 2019)

"Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation …"

- Problems: work hierarchy; responsibility; lack of knowledge (explainability); mis-match between human and machine 'stamina'

Human in the Loop in AI-Proficient: Ambiguity

- Our ethics method in AI-Proficient > specific recommendations which evolve

  (≈ 90 over 8 UCs chosen)

- E.g. UC*X* Recommendation #1 "*The default position about whether the operator x is always expected to follow the AI proposal should be specified either overall, or for various phases of AI integration if there is a trial period. A trial period with phasing in of the AI integration in stages should be implemented.*
  - *e.g. first 6 months – operator will consult AI proposals but use his own judgment whether to implement them.*
  - *next 6 months – operator must always implement AI proposals unless it is clear that AI proposal causes some major problem*
  - *Formally clarify at what stages the operator has command over the AI to the point of ignoring its suggestions if he chooses (according to the human-in-command definition)*"

History of the HITL Terms:

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- ➢ First recorded mention of HITL occurs in 1958 (Birmingham et al. Report of Naval Progress) > linked to an article of 1954 > (missile) fire control systems

  - ➢ Psychological studies; original HITL meaning: *human as component of control system*
  - ➢ Humans are too variable/adaptable, i.e. they will adapt themselves to compensate for poor system design
  - ➢ *Humans should be replaced* unless more efficient than alternative components and needed as safety monitors
  - ➢ Human as component vs. Human as monitor of system

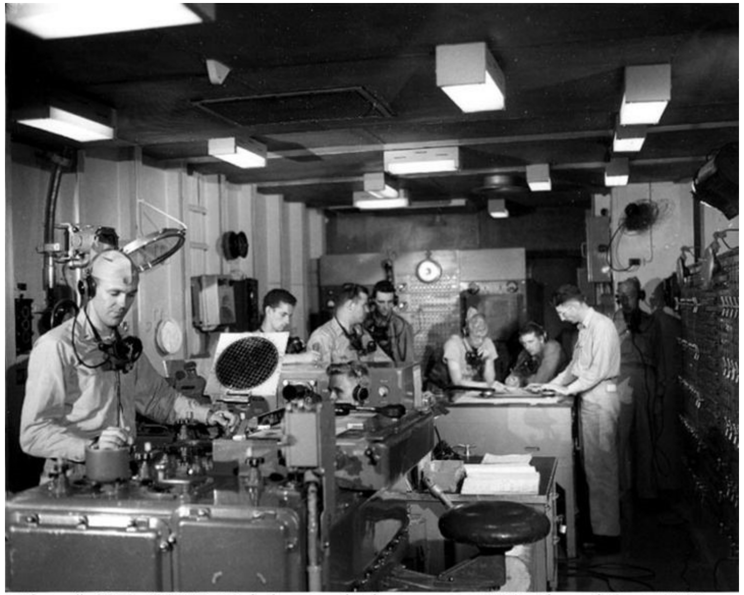


Photo # 80-G-420319 Main battery plotting room on USS Missouri, Sept. 1950

History of the HITL Terms: HOTL

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- ➢ Earliest located references from 2000 (McLaughlin et al.)
- ➢ Again the human is viewed as *a component* of control systems

- ➢ But later the term is used to indicate humans *monitoring* algorithms (drone missions)
- ➢ And also *actively guiding* automated planners for military missions

History of the HITL Terms: HOOTL

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- Earliest located references from 1988 (Sanders) > military

Human is *removed entirely* from control systems

- But again later in context of ship navigation

  term used for human being *partly removed*

History of the HITL Terms: HIC

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- Some non published EU references as early as 2017
- EU Commission Ethics Guidelines for Trustworthy AI (HLEG on AI, 2019)
  - "HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation …"

A Mess of Meanings

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- The terms have had a wide variety of meanings
- Some HITL examples:
  - Component of a system
  - Monitor of a system
  - A simulation involving real time human action
  - Human involvement *at all* in a system
  - Training an algorithm (Machine Learning)
  - + others

A Mess of Meanings

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

Term	Human Status	Publication Year(s)
Human in the Loop	Component of a control system to be simplified and replaced when possible	1954; 1959; 1976
	Monitor of a control system	1954; 1963
	Complex component of a calculating system to be replaced when possible	1979
	Hidden complex component of an automated system	1985
	Term for *a simulation* which includes real-time human action	1988
	A trainable component of a simulation	1993; 1995
	Being aided by motion control algorithms (robotics)	1998
	Being involved *at all* in the use of a computerized system	1993; 1995
	Training an algorithm (Machine Learning)	1993
	Capable of intervening in every decision cycle of a system (HLEG)	2019

Problems with the Oversight Notion

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- ➢ The newest definitions (HLEG) try to gather the terms under the notion of *Oversight*
- ➢ But oversight was not the original meaning of the terms and also abandons some of the many other meanings

- ➢ *Human as Overseer (secondary)* has been in tension with *Human as Component* notion from the beginning
- ➢ Insofar as human roles can be taken over by automation (or AI) the human is being viewed as component (Birmingham); Insofar as automation (or AI) has not encroached upon more complex human capacities the human is viewed as overseer

Ethical Implications

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- A *negative* conception of human participation
- Humans measured by how *dispensable* they are
- Inconsistent with the notion of an ethics for humans

New Terms?

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- ➢ our wish for AIs to have autonomous control is paradoxically opposed to our wish for control of the AI in terms of unwanted consequences which we nonetheless cannot foresee (Héder, 2020)
- ➢ Very visible in AI-Proficient UCs

- ➢ Oversight does not work in related areas [e.g. government use of algorithms]

(Green, 2021)

oversight is illusory!

New Terms?

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

- ➢ Oversight > transformed into *actively guiding ourselves as a community of humans engaged in technology producing actions*
- ➢ Component > transformed into *being included actively as a member of a community of humans engaged in technology producing actions*
- ➢ A positive scale, rating *human implication in the technology community*
  - ➢ *Instead of human overseeing a distinct 'object'*

IGP Rating (e.g. IGP_{111})

also described in <Marc Anderson and Karën Fort>
<Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint> (under review)

Term	Human Participation in Technology Producing Communities
Inclusion	1.*Replaceable* – technology community views a human as replaceable by any other human
	2.*Experienced* – technology community views human as experienced in multiple activities
	3.*Unique* – technology community develops technology around a human's unique skills and life context
Guidance	1.*Tester* – human's use of system is merely 'registered' (passive)
	2.*Trainer* – human's suggestions regarding system are implemented
	3.*Designer* – human designs complex parts of the system with others
Persistence	1. *Brief* – human initiates or contributes to an abstracted phase of a technology community
	2. *Sustained* – human contributes to the actions of a technology community up to the completion of the technology community's goals
	3. *Evolving* – human contributes to the process of a technology community as it overlaps and weaves into different technology communities beyond itself

An Example from AI-Proficient: UC*Y*

- Goal/project: Stabilize a problematic chemical reaction with AI
- Community: Console operators; Digital twin developers; Algorithm developers; Process engineers; Ethics Specialists, etc.
- Operator IGP ?
  - Inclusion? 1 - operator is viewed as replaceable by other operators
  - Guidance? 2 – developers will try to make use of operator feedback (suggestions)
  - Persistence? 1 – operator will implement AI recommendations eventually (but only after the design and testing phases)
- Thus ≈ IGP_{121}

An Example from AI-Proficient: Raising IGP

- Could we raise it to operator IGP$_{232}$?
  - Inclusion – could the process engineers and developers *draw upon the interests and particular skills* of individual operators? (e.g. XAI aspect of the algorithms tailored to particular operator's experience/skills)
  - Guidance – could the operators help guide the development of the solution directly (e.g. working directly with the developers in tailoring XAI, in formulating sub-phases of a solution, etc.)
  - Persistence – could we get operator input and guidance *from beginning*, e.g. in digital twin phase + *keep them informed as the AI development proceeds*?

- Oversight can still be captured in the scale for technical uses, but the oversight would then have degrees – and ethical implications would be more clear
- so IGP$_{333}$ will be true oversight (participative), while IGP$_{111}$ will be weak oversight (illusory)

AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Thank You!

This product is part of a project that has received funding from the European Union's
Horizon 2020 research and innovation program under grant agreement No 957391.