Human agency and oversight

WHITE PAPER

Trustworthy artificial intelligence

27 September 2023

Simone Kopeinik*, Sebastian Scher*, Tomislav Nad** and Dominik Kowald*

* Know-Center GmbH

** SGS Digital Trusts Services GmbH

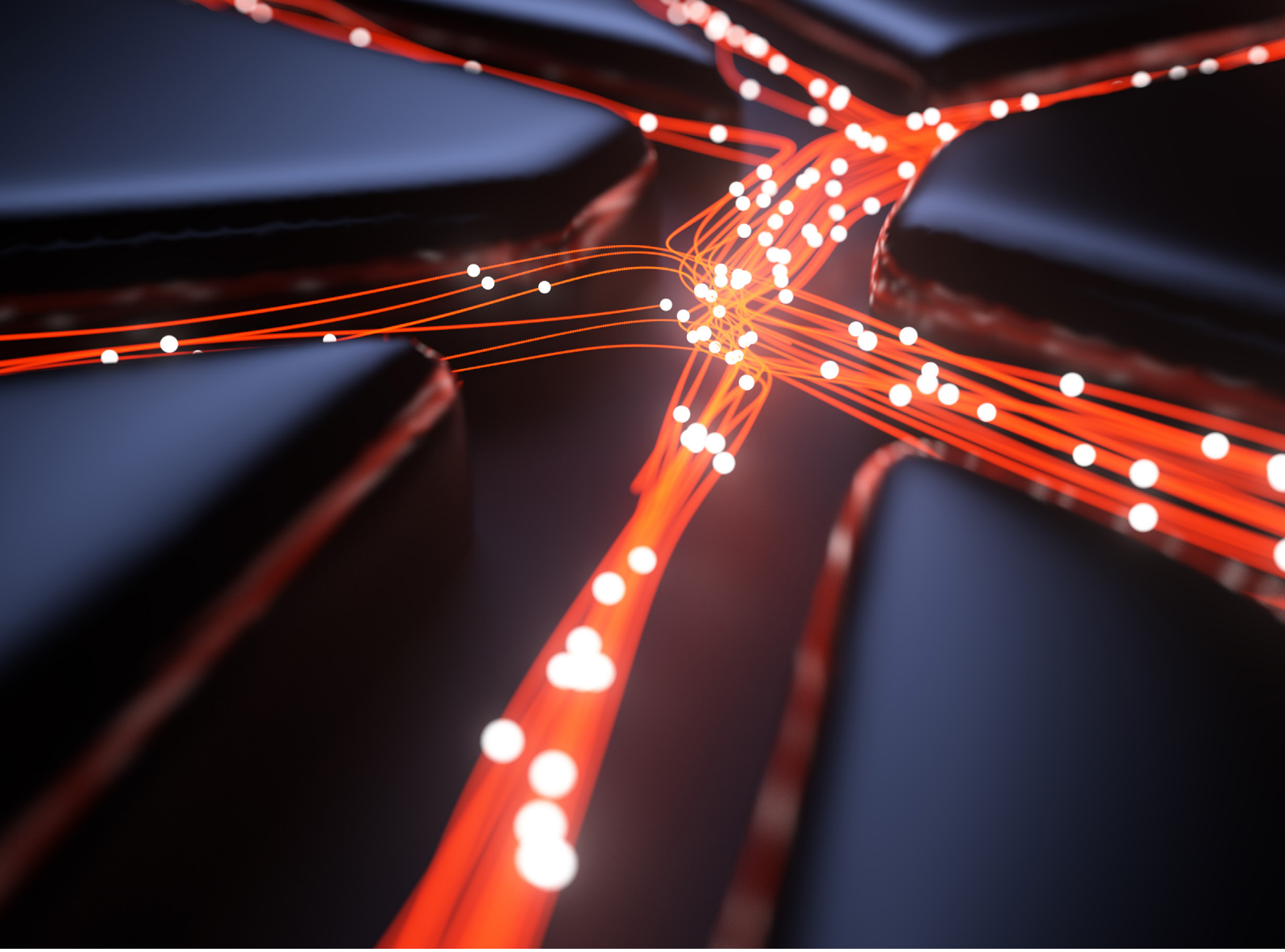
Partners of SGS

KNOW Center

SGS

Human agency and oversight

Contents:



1 Introduction

Human agency and human oversight are key principles of trustworthy AI and may ensure human-centric software and hardware systems adhere to ethical standards and fundamental user population rights[1].

The two terms are very close but are distinct from each other. Human agency in the context of AI refers to maintaining the autonomy of humans who either use AI systems, or who are exposed to the results of AI systems. While this can be understood as a philosophical concept, in practice, it also means supporting humans in conscious and informed interaction with machines; lacking manipulations, misinformation or the reductions of personal choice and freedom (e.g. caused by addiction). Moreover, humans should retain the right to intervene in automated decisions that affect them (e.g. correcting misconceptions in user models).

Human oversight, on the other hand, is directly connected with AI operations. It describes different levels of human-computer collaboration, where the human acts as a teacher or supervisor of an AI system and, thus, actively influences either the learning or the acting of the system. This encompasses activities such as monitoring, interpreting and intervening in AI operations, and requires humans to be capable and competent in the context of the AI application. In this sense, the human actor in place is supposed to minimize the risk of AI systems adversely affecting the health, security, safety or fundamental rights of the affected population. As described in the EU AI Act[2], this is considered particularly relevant in high-risk AI systems, i.e. "AI systems that negatively affect safety or fundamental rights"[3].

[1] *https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html*
[2] *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206*
[3] *https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence*

Between human agency and complete autonomy

Human agency is a concept that promotes an informed and reflective use of AI system decisions over blindly following AI output, which should enable users to pursue their goals and adhere to their values. To date, it has been only vaguely defined and often used interchangeably with human autonomy (Bennett et al. 2023). Under the umbrella of Explainable AI (XAI), technology is researched that supports human agency, providing a human-readable and understandable rationale for why a particular AI input results in an AI output. This level of transparency shall, according to (ISO, 2019, sec. 3.13), support "specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specific context of use."

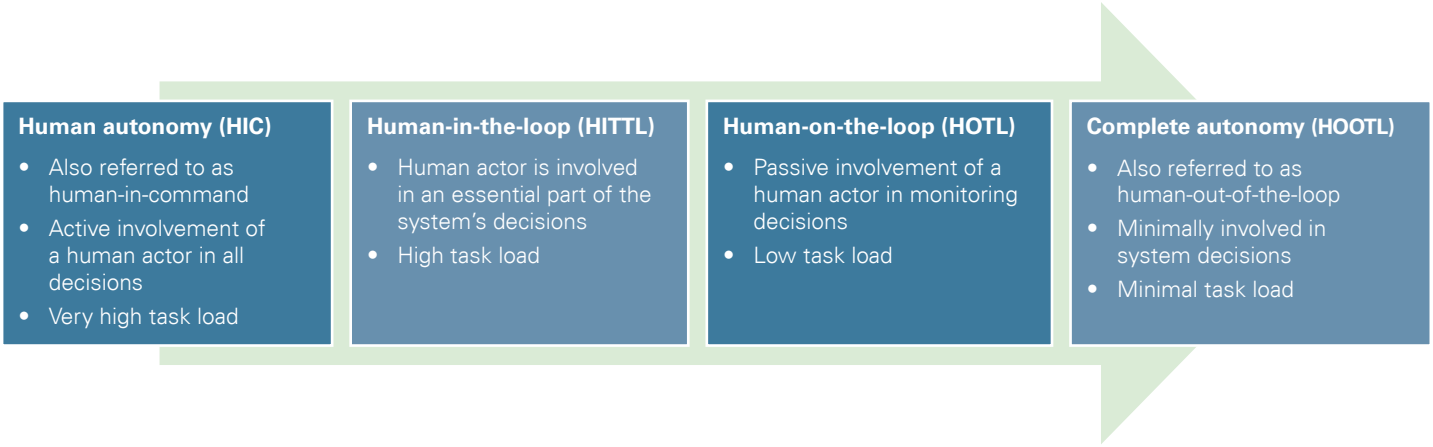
This is in line with the principle of human agency that demands human actors and users to be able to understand and consciously interact with an AI system without being subject to "unfair" manipulation or purely automated decisions. However, depending on the application context, different levels of human involvement may be sensible. This can be described by the degree of automation (Nothwang et al. 2016).

The lowest degree of automation is human autonomy. Human autonomy is defined as the operation of a system with the active involvement of a human actor in all decisions of the system. One example would be using a calculator, where the user actively inputs and triggers every operation. This causes a high level of process control and a very high task load for the human actor.

Autonomy describes the operation of a system without (constant) human input, where complete autonomy is considered the highest degree of automation, as we can, for instance, find in some deep learning applications.

Because the degree of automation required depends on the nature of the AI task and its performance goals, the approach of human-computer collaboration and the distribution of roles is selected with regard to the strength of human actors, and the strength of available algorithmic methods within the particular context. Here, the software design team of any AI system should conduct an "allocation of function" process to decide which jobs, tasks, functions, and decisions are allocated to humans and which to automation, based on factors like relevant abilities of involved humans and automation systems (e.g. speed of processing), the expected level of job satisfaction, user engagement and the desired empowerment of stakeholders.

Such a proper evaluation to determine the degree of automation is needed, as poorly thought-out task allocation can lead to a number of consequences, e.g. too low a level of human autonomy could result in mistrust, lower motivation or decoupling of the responsibilities of system users, which in turn might cause lower social acceptance and technology adoption rates (Herse et al. 2018). Too high a level of human autonomy, on the other hand, might pose risks to certain people. This mostly applies to safety-related use cases where humans are not likely to be fit to take the best decision, e.g. online applications for children which do not provide privacy restrictions or the use of heavy machinery that allows the disconnection of safety devices.



2 Human oversight: an overview of approaches for human-AI collaboration

Where human agency encompasses the broader concept of human involvement and control over AI technology, human oversight can be seen as one specific aspect of human agency, focusing on the active supervision and monitoring of AI systems to ensure they operate as intended, and to prevent undesired AI functionalities, outcomes or consequences.

The following terms categorize different approaches of human-AI collaboration (Munro, 2020, Anderson et al. 2020):

Human in command (HIC): Humans supervise the operation of the AI system and its deployment status. This approach is also discussed in terms of human autonomy, where the human actor is actively involved in all steps and decisions, and initiates AI operations on a task level. Examples for HIC encompass operating a calculator, translating a specific text in a translation app or using grammar/spelling corrections. The concept is built around the idea of humans and machines having different abilities best used in concert.

Pro	Con
Low risk of "unfair AI" causing adverse effects on society	Incorporates all kinds of human weaknesses (boredom, fatigue, anxiety, human bias)
The human agent remains in full control of the process	Very high task load on human agent

Human in the loop (HITL): Humans are an integral part (considered as a system component) of the AI process cycle. This describes a human-AI collaboration within the learning and/or decision-making processes, and is divided into two categories:

- Concepts to integrate humans into the algorithmic learning process. This category discusses mechanisms where humans either assist the machine in learning, for instance, by acting as annotators, interactive input or feedback providers, or by structuring training data, or where humans are receivers of explanations that allow them to follow and understand AI reasoning processes
- Concepts of human-AI interaction that go beyond the technical implementation of the learning process. This category addresses two main questions: i) does the AI system adhere to usability requirements during the AI learning process and the system deployment, e.g. is proper data or human expertise available (useable AI)? and ii) can the AI system be used sensibly and effectively by, and according to, the requirements of the target user group (useful AI)?

Pro	Con
High degree of transparency	Demands a fitting level of task granularity and task distribution
Utilization of human judgement	Prone to human bias
Allows for algorithmic imperfection	Not suited for strongly dynamic systems where real-time decisions are essential

Human on the loop (HOTL): Humans monitor and potentially guide the design and progress of an AI system from outside the AI process cycle. Their role is to interact with the system as a supervisor or arbiter, aiming to ensure the safe operation of the system.

In this approach, the human actor is not heavily involved, and only intervenes if the machine fails or has a high probability of failing, i.e. in the case of unexpected, uncertain events, or if crucial decisions take place.

HOTL is mostly implemented in highly dynamic systems, for instance, in manufacturing system control.

Pro	Con
Higher processing speed (no real-time human response)	Not recommended for high-risk systems unless higher safety can be achieved through automation rather than human intervention

Human out of the loop (HOOTL): Humans are not involved in the AI tasks, and the system can operate without human input. The approach is also referred to as complete autonomy or system automation, where the attempt is to operate a system without, or with only minimal, input/intervention from a human agent. The main aim is to maximize the efficiency and productivity of the system, where the technology is optimized for accuracy and physical safety. Compared to the other approaches, commercial interests are at the center of the application and are typically prioritized over user interests.

Prominent application areas of HOOTL approaches are, for example, military applications, or for vehicles or robots where AI decisions must be made in an instant.

Pro	Con
Minimal task load for human agent	Lack of human control
Comparably low costs in operation	Higher costs in development

Similarly, to the degree of automation discussed before, there is no best fit-all solution, but an evaluation of the use case and the specific application context is required to select the most appropriate approach of human involvement. Human in the loop (HITL) is the most intensively researched approach to date, which can be attributed to the high complexity of the close interconnection between humans and machines that poses challenges in various disciplines such as user interface design, pedagogics and cognitive science.

3 Open issues and challenges

The terminology of human oversight is only very vaguely defined and has been used with various meanings. These different meanings usually do not agree with each other and do not appropriately capture the complexity of human-AI interaction (Anderson & Fort, 2022). Like human oversight, human agency is also a vaguely defined concept. These concepts often overlap strongly, or in some cases, are even used synonymously (Bennett et al. 2023). Thus, the research community calls for the development of a more meaningful and clearly defined terminology that better describes human-AI interaction in all different phases of the AI life cycle.

In practical terms, finding the right balance between automation and human oversight is complex and hard to achieve. The most important factors to consider are:

- The trade-off between cost and control, as human involvement during the application lifetime is expensive
- The ethical and moral principles of the application context
- The potential impairment of human performance (e.g. fatigue, high workload, complacency) and its implications. To achieve a better understanding of these processes, extensive research on user biases when using AI systems, and its impact on the decision, as well as AI systems, is pending. This would be needed particularly in sensitive domains (Kostick-Quenet & Gerke, 2022)

Summary

Legal and ethical considerations call for the involvement of human actors in the majority of AI applications. Which approach to use and how to distribute tasks is highly dependent on context and has to be analyzed as part of the software design.

The degree of automation depends on the nature of the task and its specific requirements. Analysis of human and automation abilities is recommended for each AI system on a task level. This allows for the strengths of each agent to be best exploited. Prominent factors to be considered in this process include task type, failure modes/consequences and the autonomy-to-human error rate.

Human-AI collaboration can be categorized according to four main levels of human involvement, i.e. human-in-control, human-on-the-loop, human-in-the-loop, and human-out-of-the-loop. Most popularly researched and discussed are human-in-the-loop applications, as the extensive involvement of humans in the AI life cycle introduces a high level of complexity that is investigated by researchers from multiple disciplines (e.g. computer science, cognitive psychology and pedagogy).

Finally, the new and interdisciplinary research field around human agency and human oversight in the context of AI is still in its infancy. The existing terminology is only vaguely defined and has been used inconsistently in the past, failing to capture the complexity of the field. Further discussion and agreement on terminologies and their precise definitions, as well as interdisciplinary research on the complete AI life cycle, remain pending.

References

Anderson, M., & Fort, K. (2022). Human where? a new scale defining human involvement in technology communities from an ethical standpoint. International Review of Information Ethics.

Bennett, D., Metatla, O., Roudaut, A., & Mekler, E. D. (2023). How does HCI Understand Human Agency and Autonomy?. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-18).

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., ... & Williams, M. A. (2018). Do you trust me, blindly? Factors influencing trust towards a robot recommender system. In 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN) (pp. 7-14). IEEE.

ISO 9241-210. 2019. "Ergonomics of Human-System Interaction — Part 210: Human-Centred Design for Interactive Systems," Geneva: International Organization for Standardization.

Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. npj Digital Medicine, 5(1), 197.

Munro R (2020) Human-in-the-loop machine learning. Manning Publications, Shelter Island

Nothwang, W. D., McCourt, M. J., Robinson, R. M., Burden, S. A., & Curtis, J. W. (2016). The human should be part of the control loop?. In 2016 Resilience Week (RWS) (pp. 214-220). IEEE.

Human agency and oversight

SGS.COM/DIGITAL

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

https://know-center.at/

✉ info@know-center.at

WHEN YOU NEED TO BE SURE