Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

The Human-AI Relationship

This input paper was prepared by Hussein A. Abbass

Suggested Citation

Abbass, H. A.. (2018). The Human-AI Relationship. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

The Human–AI Relationship: Functions, Autonomy and Trust

Hussein A. Abbass

School of Engineering and Information Technology, UNSW Canberra, Australia

h.abbass@adfa.edu.au

Abstract

As artificial intelligence (AI) finds more uses in human society, a need arises to scrutinise the relationship between humans and AI. Technology itself has advanced from the mere encoding of human knowledge into a machine to designing machines that 'know how' to acquire the knowledge they need, learn from it and act independently in the environment. Fortunately, this need is not new; it has scientific grounds that I trace back to the inception of computers. The aim of this article is to share the scientific journey that many scientists over the last 50 years or more have taken to understand this relationship, and to present the nature of, and the role of trust in, the human–AI relationship. I use a risk lens to discuss risks and opportunities sitting at the human–AI interface and reveal some of the fundamental technical challenges for a trustworthy human–AI relationship. I conclude that any social integration of AI into the human social system would necessitate a form of a relationship on one level or another in society, meaning that humans will 'always' actively participate in certain decision-making loops – either in-the-loop or on-the-loop – that will influence the operations of AI, regardless of how sophisticated it is.

1 Augmenting Human Functions with Technological Functions

Since the inception of the human race, technologies have been an integral part of human society. The Oldowan stone tools (Susman, 1991) assisted humans 2.6 million years ago in farming, hunting and construction. Simple tools created opportunities for humans, providing them with more resources to support their families and enabling them to improve their quality of life. For millions of years, these technologies physically augmented the human. Only a thousand years ago, when the Chinese invented Suanpan (Flegg, 1989; one of the first forms of a calculator known in history), did humankind start to see the birth of cognitive augmentation: tools that help humans to think faster and do complex counting and arithmetic operations that cannot be done by the human brain alone. This form of augmentation enabled humans to be more efficient in trade.

Digital calculators then added greater functionalities that ranged from complex calculations to the ability of storing and memorising information. As humanity started to develop the first electronic digital computers through Babbage's work in 1821 (Randell, 2013), new extensions to human cognition began to appear. Present day technology enables a computer to augment humans' planning abilities, as in the case of a GPS planner in a car, and can memorise items to extend human memory, as in the case of memorising appointments using a calendar. Sensor technologies with natural biological sensors enable humans to see, hear and feel things they could not see, hear or feel before. Robotic actuators allow humans to extend their body (Controzzi, Cipriani, & Carrozza, 2014; Goh et al., in press) when they lose an arm or a leg, or when they need extra strength, by using an exoskeleton to carry more weight than they could normally support. The technological landscape has evolved steadily from simple automation to advanced automation that can respond better than a human in a specific situation. For example, a collision avoidance system in a car can sense danger faster than a human and can execute a resolution strategy in a time critical, life-or-death situation. Despite this long history of physical and basic-level cognitive augmentation, never before have humans been able to augment their intelligence. Tools, including computers, have always been under significant human control. The human is the master and the tool has always been the slave that has no mind of its own to challenge its human master.

Artificial intelligence (AI) promised for many years to revolutionise this form of augmentation. Since its inception, AI has promised to solve problems on behalf of the human independently; it can understand humans and communicate with them, and it can even challenge humans in their unique characteristic: natural intelligence. The Turing test (Turing, 1950) has been the primary test for AI, challenging AI developers to design an AI to be indistinguishable from humans. However, for decades, the promise of AI was bigger than its reality, creating valleys of AI death.

Over the last decade, the fate of AI has started to turn, together with its enabling technologies, such as sensors, communication, the Internet, computer speed and storage. Micro-services (Smallegange, Bastiaansen, Venema, & Bronkhorst, 2018), which break down large tasks into many small programs that can be distributed everywhere and anywhere and be summoned on demand when the need arises, offered industry an opportunity to replace a single giant AI program with many independent services performing specialised functions (Dragoni et al., 2017). This concept of micro-services is used today by many companies, including Microsoft, and has been the basis for some of the most recent media

stunts, such as those associated with the robot Sophia becoming the first robot to be granted citizenship by a country (Abbass, 2017). Micro-services will not only conceal the reality of AI affairs today but will also allow sudden unanticipated tipping points to appear in the technological landscape of AI. As we approach these tipping points, we need to pose the question: what is AI exactly, and what roles should humans play in safeguarding society against AI?

2 Artificial Intelligence – A Technological Definition

There are many definitions of AI (Bringsjord & Schimanski, 2003; Fetzer, 1990), thanks to the complexity of expressing the concept of intelligence in finite words and the blend of beauty and ambiguity in human language. These attempts spent little time arguing what 'artificial' refers to in AI and much time arguing what 'intelligence' is. Properly, computational intelligence is a more adequate terminology to avoid the philosophy of what artificial is and what it is not, and to focus more technologically on the fact that the AI I discuss in this article is of a computational nature. Nevertheless, I start by offering my own definitions for AI for two reasons. First, to communicate to the reader what AI means to me, which will facilitate an understanding of the remainder of this paper. Second, to offer a structure for this paper that naturally unfolds the relationship between AI and humans as well as that between AI and other concepts and research areas, such as autonomy, smart autonomous systems, trusted autonomy and robotics.

I begin with my functional definition of AI:

> *Artificial intelligence is concerned with the design of computer algorithms, methods and methodologies that enable machines to: understand the world and assess themselves and their context to identify hazards, threats and opportunities affecting their goals; generate, choose and execute appropriate courses of actions to achieve their goals; learn to improve their performance and adapt to changes in their surroundings; and educate and transfer their knowledge to others (humans and machines).*

I like to simplify the above definition technologically to the following:

> *Artificial intelligence aims to design algorithms to provide computers with cognitive skills and competencies for sense-making and decision-making.*

These two definitions are anchored in the underlying philosophy of this paper. The first lists the characteristics of different AI algorithms to be:

- the ability to interpret data, represent and understand context and situations
- the ability to assess opportunities and risks in contexts and situations
- the ability to generate courses of actions, select and execute one or more of them
- the ability to learn and adapt
- the ability to share knowledge by transferring it to other AI's or, through explanation, to a human.

The second definition posits AI as the automation of cognition (machine cognition) to develop skills and competencies to perform tasks. It highlights the two major streams of applications we see in today's world of AI: data analytics, which focuses on analysing, interpreting and transforming data into knowledge, and autonomy, which focuses on producing actions that assist the AI in achieving its own design objectives.

In data analytics, a core step for sense-making, the results of the analysis inform humans or another AI to identify an appropriate course of action. Since data analytics does not produce decisions per se, the risk of that AI needs to be managed by those who will use the output of data analytics to generate the actions. For example, a panel display in a vehicle shows the result of the analysis performed by the AI to transform raw signals received by the car into information to be used by the driver, for example, external temperature information and navigation. The human could choose to ignore this information or use it as appropriate.

Decision-making produces actions based on the data and understanding provided to the agent (a human or an AI). If the decision-making agent acts directly on the environment, part of the risk of that decision is transferred to the inputs to that agent. A loan assessment tool makes decisions by itself, and acts on the information it receives. If the tool guarantees to make the right decision, this would likely be conditional on receiving the correct information with an appropriate level of accuracy. This tool could still make the wrong decision if the human feeds it with incorrect information. Nevertheless, the tool might simply make a recommendation to a human who has the authority to accept or reject this recommendation; hence, the risk of the decision could still be managed.

Autonomy requires an AI with both sense-making and decision-making abilities, as well as the ability and authority to execute the decision. This form of AI senses information from its environment, assesses context, makes decisions, executes the decisions and is authorised to execute these decisions. An autonomous loan agent would collect information about the

borrower, analyse the risk profile of the borrower and their financial abilities, decide the size of loan the borrower may obtain, initiate the loan in the system and authorise the loan.

When an autonomous loan agent works together with a human financial adviser, the human–AI relationship needs to be considered at the design stage of the AI development and during human training. Since the relationship has been a long-studied research topic, it is pertinent to discuss the literature for the reader to appreciate that the design of this relationship could happen on a fine level of granularity, which manages the risk for the society. While challenges exist, as I discuss later in this article, an important piece of enabling science revolves around the concept of function allocation.

3 Function Allocation to Humans and Machines

At the design stage of a new technology, the potential missions and contexts for which the technology will be used are analysed to identify the different functions required for these missions. If the mission is to drive a car from one place to another, we could have a function to observe the environment, a function to evaluate the observations to understand the current location of the car, a function to detect hazards, a function that plans the next location of the car, a function to decide on appropriate acceleration and de-acceleration rates to control the velocity of the car, and a function to steer the wheel. These seem sufficient for our purpose to avoid unnecessarily complete enumeration of this cumbersome task. When a car has some level of automation, which functions should we delegate to automation, when and how? These questions have been the subject of a long history of research focusing on function allocation: the process by which functions are allocated to humans and machines. It is a form of division of labour. Function allocation raises three key questions:

- Methodology: How should functions between humans and machines be allocated?
- Responsibility: Who is doing the allocation?
- Authority: Who can authorise an allocation?

3.1 Function Allocation Methodologies

Two distinct categories of methods exist for function allocation: static and adaptive allocation. Rouse and Rouse (1991) defined three classes of static allocation:

- comparison allocation, where the better performer is chosen; that is, if the human is better than the machine in one function, the human is chosen, otherwise the machine
- leftover allocation, where every function that could be automated is allocated to automation and only those functions for which no automation is possible are allocated to humans
- economic allocation, which uses a cost–benefit analysis approach, where if automating a function is not cost-effective, even if it could be automated, it is assigned to a human.

Static allocation presents a multitude of problems. First, it assumes that the suitability of a human or a machine for a function does not change over the course of the mission. This is clearly not the case, because the state of the environment, AI and humans are all factors that influence the suitability of a human or a machine for a particular function in a particular context. In a situation with severe consequences, a function that is performed by a machine might need to be performed by a human, and vice versa – a function that a human does well under normal workload conditions might need to be switched to a machine if the human is overloaded.

Second, none of the three classes of static allocation described above considers the human element. Each sees humans as machines, and the allocation is based purely on factors such as performance and cost. This could lead to assigning humans uninteresting functions, causing human boredom and demotivation. A third problem is the underlying assumption that it is the designer who decides what to allocate to humans and machines. The human operator does not have a say and is assumed to 'listen' to whatever the designer has decided when using this form of 'technologically centred design'.

The problems described above gave birth to a new category of function allocation called adaptive allocation, developed to serve a 'human-centred design'. Different names of this category emerged, including adaptive aiding, adaptive automation, and adaptive allocation, but the origin can be traced back to the early 1960s (Babcock, 1960; Berry, 1961). The concept was popularised with a United States Air Force project in the 1970s on cockpit automation (Pope, Bogart, & Bartolome, 1995).

Adaptive allocation changes function allocation during a mission. Rouse (1994) discussed three strategies (called automation allocation logic [AAL]) to control *when* a change in function allocation needs to be triggered:

- critical event logic, where automation hands over a function to a human if there is a critical event or vice versa
- measurement-based logic, where the handover could occur in any direction at any point in time based on a continuous measurement of a state such as human workload level

- modelling-based logic, where human performance is captured in a model that predicts future human workload and makes decisions about when a different function allocation is needed.

The earlier versions of adaptive allocation focused on simple models of human performance that were built through observational research. This was true for adaptive aiding (Berry, 1961) and adaptive automation (Babcock, 1960). Research areas such as human–machine interaction (Fitts et al., 1991), man–machine symbiosis[1] (Licklider, 1960), human–machine teaming (Susman, 1991) and human–autonomy teaming (Sheridan, 2012) have largely relied on simple models to either analyse or model the relationship. Meanwhile, a second evolutionary path has been growing steadily since the 1970s, which has led to the newer concepts of augmented cognition (Schmorrow, Stanney, Wilson, & Young, 2006) and cognitive-cyber symbiosis (CoCyS; Abbass, Petraki, Merrick, Harvey, & Barlow, 2016).

However, dynamic function allocation could distract humans, disturb their situational awareness and, consequently, negatively affect safety and performance in the tasks assigned to them. Moreover, if the allocation is based on subjective data and/or objective data that interfere with the task with which a human is involved, the allocation errors could increase, resulting in higher errors in the human–AI relationship. Therefore, the use of augmented cognition and CoCyS attempts to overcome these limitations.

Augmented cognition (Schmorrow & Kruse, 2002; Schmorrow et al., 2006) is encephalography (EEG)-based adaptive allocation. EEG represents the signals occurring because of brain activities that are triggered by brain functions. In 1875, Richard Caton demonstrated that fluctuations in brain activities follow mental activities. This finding meant that brain signals could possibly indicate the impact of the mental processing a human is experiencing while performing certain functions. However, Caton's work (as cited in Demos, 2005) was demonstrated on animals. It was not until 1929 that Hans Berger (1929) published the first study to record EEG data from humans. In 1934, Adrian and Matthews (1934) demonstrated that not only are brain waves triggered by mental activities but also by external stimuli, which could influence the way they are triggered. This idea was possibly the first scientific evidence for neurofeedback, a recent field in clinical psychology (Demos, 2005), which retrains the human brain to produce signals to correct physiological causes of some psychological illnesses, such as attention deficit disorder.

In 1970, the United States Department of Defense (DoD) recognised the potential of this maturing line of research and embarked on two projects on biocybernetics (Hill, 1973) and brain computer interfaces (Vidal, 1973). The ability to record and interpret human EEG opened a variety of opportunities, including the potential for deriving workload, attention and situational awareness objective indicators from EEG to guide AAL, even in real time in an operating environment. Another United States DoD activity in the early 2000s led to the new concept of augmented cognition. As EEG sensors developed and became more reliable, the science evolved to a reasonable technological maturity level, and function allocation research using EEG became well established. In this line of research, function allocation logic has possibly been the simplest form of AI. It was pre-designed, lacked flexibility because it did not change its behaviour once deployed and, simply put, lacked the smartness and autonomy to match the complexity of the AI it was attempting to manage in the first instance. This led to the Australian born CoCyS (Abbass et al., 2916) concept, which revolutionised the adaptation process from a machine adapting to a human to smart adaptive agents (called 'e-cookies') that act as autonomous relationship managers between humans and machines.

Each e-cookie analyses different data sources (e.g. EEG, muscle movements, heart rate, skin temperature, eye tracking, voice, images) measured from the human under observation, objectively inferring human states (e.g. workload, attention, emotion). It equally analyses different data sources from the automation itself to understand autonomously the state of the task and predict its future states. An e-cookie then learns from this information to identify when and how to adapt the function allocation strategy. An e-cookie also acts as a smart regulator and/or smart safety net to ensure the trustworthiness of the relationship.

CoCyS opened the way to a new and more sophisticated form of AAL, one that is trust based, allowing an e-cookie to reallocate the functions between humans and machines based on trust indicators. If a machine is hacked or spoofed, some functions are retracted from the machine and allocated to the human. If there is a suspicious change in human identity, some functions that were performed by the human are reallocated to the machine.

Adaptive allocation created a control agent between humans and machines. To avoid confusion between the automation we have discussed so far and the automation of the agent responsible for adaptive allocation, I refer to the latter as the allocation agent. An example of an AI-based allocation agent I have discussed above is e-cookie. This naturally raises the question of who is doing the allocation agent job, and who has the authority to make a decision to execute the allocation agent's recommended course of action. These two questions will be addressed next.

[1] The word 'man' in Licklider's language is not necessarily gender biased because an older use of the word in English had a meaning of any 'human'. Most of today's reference to the concept would use human–machine symbiosis.

3.2 Function Allocation Responsibility

Humans can control the allocation agent; thus, they can assess a situation and adopt a strategy of when they wish to push a function to the AI and when they wish to retract a function from the AI. Equally, the control of the allocation agent can be automated. For example, when the system notices a critical event, it can reallocate a function from a human to a machine or vice versa based on a set of predefined rules. E-cookie is an example of advanced automation of the allocation agent. However, what type of intervention triggers are needed for the human or AI to intervene, when and how? This line of thinking, however, depends on the level of automation and the role the human plays in the system.

Levels of automation (LOA) refers to the maturity level of automation. Sheridan (1992) and Sheridan and Verplank (1978) defined 10 levels using a 'who-centred' approach, while Endsley and Kaber (1999) defined another 10 levels from an information-sharing and situation awareness perspective. The latter were then refined, with Endsley (2017, 2018) increasing the levels to 12. The two approaches are presented below:

Endsley's Levels	Sheridan's Levels
L1: *Manual Control*: Human performs all aspects of tasks. L2: *Information Cueing*: Computer aids by highlighting key information on screen or decluttering irrelevant information. L3: *Situation Awareness Support*: System gathers key information and integrates for levels 2 & 3 of situation awareness. L4: *Action Support / Teleoperation*: Computer aids by executing each action as instructed. L5: *Batch Processing*: Computer completely performs singular or sets of tasks commanded by human. L6: *Shared Control*: Computer and human generate decision options; human decides and performs with support. L7: *Decision Support*: Computer generates recommended options, human decides (or inputs own choice) and system performs. L8: *Blended Decision-Making* (Management by Consent): Computer generates recommended options and selects best; human must consent (or override) and system performs. L9: *Rigid System*: Computer generates recommended options from which human may select (cannot override) and system performs. L10: *Automated Decision-Making*: Computer generates recommended options along with human; system selects best and system performs. L11: *Supervisory Control* (Management by Exception): Computer generates recommended options, selects best and system performs; human can intervene if desired. L12: *Full Automation*: Computer performs all aspects of task with no human intervention possible.	L1: Human performs whole job up to the point of turning it over to the computer to implement. L2: Computer helps by determining the options. L3: Computer helps to determine options and suggests one, which human need not follow. L4: Computer selects action and human may or may not perform it. L5: Computer selects action and implements it if human approves. L6: Computer selects action and informs human in plenty of time to stop it. L7: Computer performs whole job and necessarily tells human what it did. L8: Computer performs whole job and tells human what it did only if human explicitly asks. L9: Computer performs whole job and decides what the human should be told. L10: Computer performs the whole job if it decides it should be done and, if so, tells human, if it decides that the human should be told.

LOA have evolved in a culture of decision-making in safety critical systems where the physical body of the AI in the form of a robot is not part of the operator's responsibility and, therefore, is irrelevant to the relationship between humans and machines. Nonetheless, in the field of human–robot interaction, the relationship between the human and the robot can take different forms.

Scholtz (2003) defined five roles for humans in human–robot interaction: supervisor, operator, teammate, bystander and mechanic. In a supervisory role, the human oversees what the machine does and advises accordingly. A supervisor does not become involved with low-level tasks for which the machine is responsible. Instead, the supervisor takes meta-actions, and makes plans and high-level decisions. An operator, however, monitors low-level action-level tasks. A teleoperation scenario is an example of a human operator responsible for performing low-level tasks. A teammate is a role in which the human works collaboratively with the machine to perform the mission. They can both issue advice to each other and delegate tasks to one another. As a mechanic, the human modifies any abnormal behaviour the machine displays or fixes a mechanical problem with the machine. A bystander acts as a facilitator between the robot and the environment, for example, a bystander might remove an obstacle from the robot's path if the robot is not designed for managing obstacles or if the robot is so expensive and fragile that such damage would be costly.

In summary, a supervisor *guides* the robots, an operator *controls* low-level actions, a teammate *collaborates* with the robot to perform the mission, a mechanic *fixes* the robot or its mistakes, and a bystander *modulates* the relationship between the

robot and the environment. Other forms of human–machine relationships can be seen in the LOA table above, such as shared control, supervisory control, mixed initiatives or blended decision-making.

These different forms of human–AI relationships have a significant impact on function allocation. A static function allocation works in simple tasks where the exact division of labour between the AI and the human is clear. In more complex tasks, where the environment is naturally uncertain, it becomes more difficult to follow a static function allocation. The authority for function allocation becomes more important than ever and should be considered in the design of the AI. Who should authorise a reallocation? Should the function allocation agent itself be allowed to change? These questions will be addressed next.

3.3 Function Allocation Authority

The allocation agent could be a human or an AI. It might be delegated to authorise a change in function allocation or another human or an AI might authorise the change. In the latter case, the allocation agent makes a recommendation, while the authoriser accepts or rejects this recommendation.

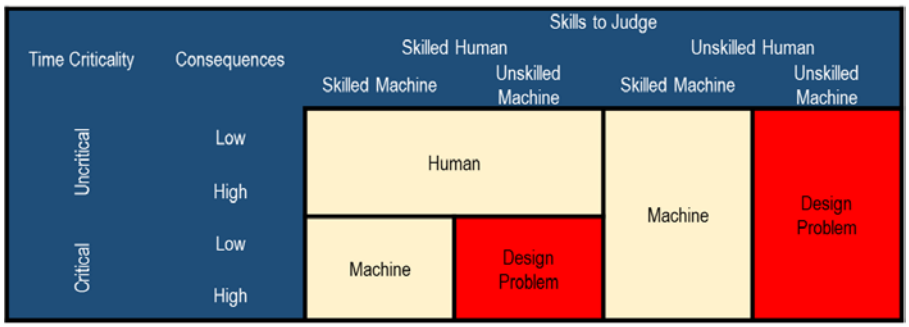
Regardless of the nature of the allocation agent, the question remains: what are the conditions for a machine to authorise a change in function allocation? Inagaki (2003) offered examples for situations that could justify the machine authorising such a decision and noted:

> *The automation may be given the right to take an automatic action for maintaining system safety, even when an explicit directive may not have been given by an operator at that moment, providing the operators have a clear understanding about the circumstances and corresponding actions which will be taken by the automation.*

While safety is indeed an important factor, it seems more appropriate to generalise Inagaki's perspective using a risk lens. I distil from the literature three factors that should influence this decision: time criticality, skills to judge and severity of consequences. In a situation where the time needed to act is insufficient for a human to make the decision, either the machine needs to decide or the consequences of inaction need to be evaluated.

A function is time critical when the difference between the time by which the decision needs to be made to be an effective decision and the time required to make that decision becomes smaller and approaches zero. An allocation agent in an aircraft flying using the autopilot system might foresee a weather cell five minutes ahead. The allocation agent assesses that the complexity of this situation is not suitable for the autopilot to fly the aircraft and hands over control to a human. The five-minute window allows the human to overcome the effect of a surprise, comprehend the situation and prepare for action. If the time window were five seconds instead, this would not provide sufficient time for the human to perform the function. Therefore, time criticality depends not only on the decision to be made but also on the capabilities of the agent that will make the decision, the context and the exact situation the agent is facing.

The second influencing factor, skills to judge, relates to when the human or machine is not sufficiently skilled to authorise; that is, the human or machine lacks the skill to judge whether a change in function allocation should take place or not. The third factor concerns whether the severity of consequences is high or not. The table below (or Skills to Judge) summarises all possible combinations of these factors and the recommended authority for the decision.



The red areas highlight the need and importance of the pre-design analysis to ensure that either the human or the machine has the necessary skills to judge the appropriateness of a change in function allocation.

The table could be further adopted to suit the specific context for adaptation to occur. For example, while a human could authorise the action when the time is uncritical and even when the machine is skilled, a skilled machine could equally authorise the action if the human is overloaded.

4 Relationships of Equals: Why is Teaming Hard for AI Agents?

The discussion so far has highlighted the function allocation problem and described the relationship between humans and machines purely in terms of function allocation. As the LOA for machines approach Level 12, the relationship between humans and machines changes in nature. Function allocation focuses the effort on well-engineered concepts, leaving out elements where machines could have a different intent from humans and could even 'supervise' the human. Take, for example, a robot teacher supervising a child while learning multiplication, or a robot coach supervising a group of swimmers.

In these cases, current LOA are insufficient, and therefore limited in their abilities to describe these future AI systems; hence, it would be restrictive to limit the discussion of challenges in the human–AI relationship using only LOA.

Scholtz's (2003) roles, discussed above in Section 4.2, could be useful here; they can be summarised functionally as to *guide*, to *control*, to *collaborate*, to *fix* and to *modulate*. With smart (advanced AI-based) autonomous systems, we should assume that these functions are available for both humans and automation. In one situation, the human might supervise a robot performing low-level control, while in another, the human might perform the low-level control task while the robot supervises the human. A first-aid robot could perform a mechanical role to treat humans if they are injured.

When the AI takes an equal role to a human, the level of control could take a multitude of forms, as shown below:

	Human-led	AI-led
Guidance	Human guides, AI performs: a surgeon guiding an AI during a medical procedure (Taylor, Menciassi, Fichtinger, Fiorini, & Dario, 2016).	AI guides, human performs: AI guiding a pilot to control the workload of air traffic controllers Abbass, Tang, Amin, Ellejmi, & Kirby, 2014).
Control	Human controls, AI senses and actuates: human teleoperating a vehicle in mining (Hainsworth, 2001).	AI controls, human senses and actuates: a smart intelligent artificial limb.
Collaboration	Human and AI work together, possibly as equals: collaborative planning (Klien, Woods, Bradshaw, Hoffman, & Feltovich 2004).	
Fixing Mechanic	Human fixes, AI performs: a human engineer adjusting the actuators of a humanoid robot to avoid falling while walking.	AI fixes, human performs: an AI modulating the level of insulin in human blood (Kovatchev, 2017).
Modulation By-standing	Human modulates the environment, AI does: human removing obstacles from the path of a robot that could damage the robot.	AI modulates the environment, human performs: an AI clearing land mines for humans to advance (Reddy, 2006).

Clearly, before AI takes more control in the world, it should reach a level of maturity that makes it, at least, able to collaborate with humans. Human–AI collaboration is a non-trivial problem. Klien et al. (2004) listed 10 different challenges that the technology faces before AI can reach the level of maturity required to collaborate with humans as equals. I list these challenges here for completeness:

Ten Challenges of Human-Automation Collaboration, as listed in Klien et al. (2004)
Challenge 1: To be a team player, an intelligent agent must fulfil the requirements of a Basic Compact [a commitment of goal alignment] to engage in common-grounding activities.
Challenge 2: To be an effective team player, intelligent agents must be able to adequately model the other participants' intentions and actions vis-à-vis the joint activity's state and evolution – for example, are they having trouble? Are they on a standard path proceeding smoothly? What impasses have arisen? How have others adapted to disruptions to the plan?
Challenge 3: Human–agent team members must be mutually predictable.
Challenge 4: Agents must be directable.
Challenge 5: Agents must be able to make pertinent aspects of their status and intentions obvious to their teammates.
Challenge 6: Agents must be able to observe and interpret pertinent signals of status and intentions.
Challenge 7: Agents must be able to engage in goal negotiation.
Challenge 8: Support technologies for planning and autonomy must enable a collaborative approach.
Challenge 9: Agents must be able to participate in managing attention.
Challenge 10: All team members must help control the costs of coordinated activity.

Scrutinising the 10 challenges above, one can easily observe the gap between AI technology as it stands today and the requirement for AI to be able to collaborate as equal to humans. For example, the second challenge calls for AI to be able to model humans' intentions and humans to be able to model AI's intentions. A significant amount of research is still under way to infer human intention in *simple* human–robot interaction tasks with unsolved problems, never mind complex interaction tasks. The research community is still developing solutions for AI to be transparent and explainable to allow humans to understand the intention of an AI and develop mutual predictability and shared understanding.

It is tempting for some to claim that taking the human out of the loop will make collaboration easier because different AIs could exchange intentions more efficiently in their own computer language than by attempting to learn intentions or communicate them with a human. Even in this situation, challenges 1, 7, 8, 9 and 10 impose significant burdens on the current most-advanced AI systems. For example, negotiation alone is a computationally expensive, mostly intractable problem (Dunne, Wooldridge, & Laurence, 2005).

The discussion above demonstrates that the development of an AI that could truly collaborate, as an equal teammate, with humans is technologically distant from today. Even if an AI exists that is so sophisticated that it can claim to be more intelligent than a human, such an AI will struggle significantly to work as a team member with other AIs because of the computational complexity required to scale the AI for team negotiation and self-synchronisation of actions. Most of the challenges above could have solutions once the context has been appropriately limited. The technological reality discussed above for AIs working in open-ended contexts does not preclude us having in the near future very advanced AI systems that could truly outperform humans in specific real-world tasks. These advanced AI systems will require effective methodologies for their social integration in the human system and a level of trust to allow them autonomy in their specific operating context. These AI systems form the context for discussion in the next section.

5 Human–AI Trust

Trust is the glue of social systems because it assists humans to manage and reduce complexity in the world (Luhmann, 1979). Some of the factors that allow a trustor to trust a trustee include the ability to carry out a given task, benevolence towards the trustor, and integrity such as fairness and honesty (Mayer, Davis, & Schoorman, 1995). Kim (1967) noted listeners' perceptions of a speaker's expertness, reliability, intentions, activeness, personal attractiveness and the majority opinion of the listener's associates as elements of trust. Thus, trust blends a complex array of interaction factors including attitude, beliefs, control, emotion, risk and power.

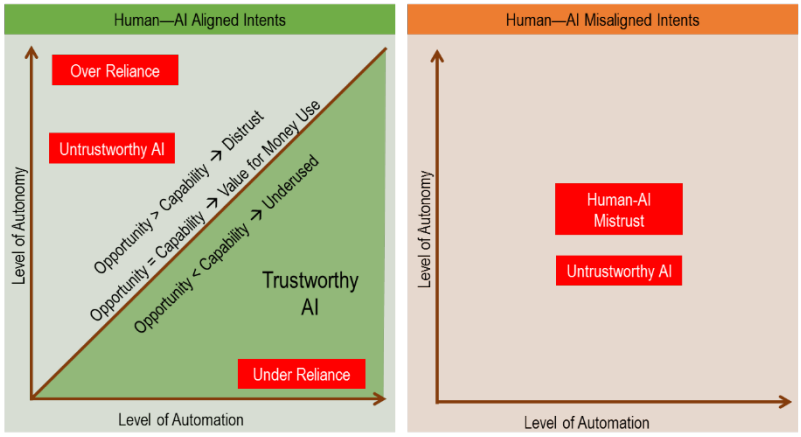
Behavioural psychologists see trust from a social dilemma lens. Deutsch (1962, 1973), for example, viewed trust as a path of ambiguity that could lead to one of two possible outcomes: one is perceived to have negative valency that is greater than the positive valency perceived to be associated with the second. The trustee controls the outcome and decides which event will occur. If the trustor chooses this path, the trustor is said to trust the trustee; otherwise, the trustor distrusts the trustee.

The literature on trust is immense, with definitions from many perspectives. Two common definitions are selected. Mayer et al. (1995) defined trust as 'willingness to be vulnerable to another based on the expectation of favourable outcomes for the trusting party'. Lee and See (2004) defined trust as 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability'. The common thread in these two definitions, and many other definitions in the literature, is that trust exposes a person to vulnerability (CoCyS; Abbass, Petraki, Merrick, Harvey, & Barlow, 2016). Borrowing definitions of vulnerabilities from the risk assessment literature could help us better understand human–AI trust. A vulnerability could be broken down as follows (Abbass, 2015):

$$\text{Vulnerability} = f(\text{Capability}, \text{Opportunity}, \text{Intent})$$

The above relationship explains the three dimensions affecting a vulnerability. The first is the capability of the trustee. The more capable a trustee is, the more serious a vulnerability could become because it could cause more damage if the trustee defects. The second is the opportunity that the trustor gives to the trustee. When the trustor trusts the trustee, the trustor gives the trustee an opportunity to defect. The third is the intent of the trustee. A capable trustee that is given an opportunity (is trusted) will not defect if the intent is good.

The above explanation fits perfectly with AI. The capability of an AI represents its skills and competency levels, and thus, the level of automation of that AI. The opportunity is the level of autonomy, the degrees of freedom with which the AI is permitted to execute its actions and the authority delegated to it. The intent in simple AI agents is the design intent, whereas the intent in AI agents that learn and adapt could change as they interact with the environment. Thus, human–AI trust could be mapped across the two dimensions of level of automation and level of autonomy, assuming the AI's intent is aligned with the human's intent.



The figure summarises the human–AI relationship using a vulnerability lens. If the

Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

The Human-AI Relationship

This input paper was prepared by Hussein A. Abbass

Suggested Citation

Abbass, H. A. (2018). The Human-AI Relationship. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

intent of the AI is not aligned with that of the human, the AI is likely to make decisions that disappoint the human and cause the human to suspect the intent of the AI, thus leading to a situation of human mistrust regardless of the AI's level of automation (capability) and level of autonomy (opportunity). If the intent of the AI is aligned with the human, automation and autonomy start to moderate the relationship. When the level of automation is lower than the level of autonomy, the AI is given opportunities that exceed its abilities, leading to the human over-relying on the AI, culminating in disappointment, distrust and an untrustworthy AI. When the level of automation is greater than the level of autonomy, the AI is more capable than the opportunities it is given. In this underutilisation case, the trustworthy AI performs the tasks successfully despite being underused but the high cost of producing this AI is not justified with its use. When the level of automation matches the level of autonomy, the AI is trustworthy, and the overall system is balanced.

The risk in the human–AI relationship could be traced to who is doing what in each of the four components (sense-making, decision-making, execution ability and execution authority), resulting in the following possible situations:

Human Control	Sense-Making	Decision-Making	Execution Ability	Execution Authority	Nature of Risk
Absolute	H	H	H	H	Limited human cognition and bounded rationality could lead to high errors, information overload, and inability to manage complex tasks.
High	AI	H	H	H	Undesirably biased analytics could drive the human to unfair decisions, while human bias and limited cognition could add more complexity to the mix.
High	H	AI	H	H	Undesirably biased recommendations could make the human accountable for unethical or legally uncompliant decisions, although the human could be overwhelmed by the available data, and their own bias and limited cognition could add more complexity to the mix.
Medium	AI	AI	H	H	In the absence of transparency and explainability of the AI, the human does not have enough information to form a judgement regarding the chosen decision. Information and situation complexity could overload the human. The human could become accountable for inappropriate decisions.
Low	AI	AI	AI	H	In the absence of transparency and explainability of the AI, the human has no understanding of the rationale of the decision. Information and situation complexity could overload the human. The human's accountability is blinded.
Low	AI	AI	H	AI	The AI controls human actions and could lead the human to wrong actions.
None	AI	AI	AI	AI	The human is out of the loop, legal responsibilities and accountabilities of the decision are both unclear.

The above matrix assumes that a full block is assigned to either the AI or the human. For example, the overall sense-making functional-block is a human's responsibility alone or the AI's alone. The human–AI relationship is normally designed on a finer level of granularity. Since the interaction of a human with an AI would invoke multiple functions, Lee and See (2004) referred to this fine level of granularity as function specificity, where the human interacts with specific functions. One function may be trustworthy, while another may not. This situation may cause the human to distrust the overall system. This naturally takes us back to function allocation and the need for a smart allocation logic, such as e-cookie, to manage the relationship and ensure trustworthy human–AI interaction.

6 Many Questions, Few Answers

The paper has drawn on a large transdisciplinary body of literature to demonstrate that there has been a wealth of research conducted in which scientists, technologists and engineers have thought about the relationship between humans and AI. While methodologies exist, new fundamental and challenging questions continue to emerge. The complexity of AI is increasing; thus, what used to work just in a structured safety critical environment, such as air traffic control, needs now to be used in unstructured environments, in the hands of the public, and in situations where it is not necessarily possible to think through responses in advance. This leaves us with more questions than the current state of science can readily answer. I will focus on a few.

Should humans have control over AI? AI operates in a social context. Whatever roles it will play as a technology, it will be serving a role in society. Some aspects of this role are better performed by AI in which human intervention could be undesirable, while other aspects need human approval. For example, a self-driving car should self-manage itself when an obstacle suddenly appears. In this situation, any delegation to a human could be fatal, since the human does not have the cognitive capacity to respond in time. The destination of the self-driving car is a human choice, which the human needs to approve. The answer, therefore, is not a straightforward yes or no; it is a matter of when and how. Function allocation gives us the scientific foundations to search for an answer. As more AI becomes integrated into society, we need to simultaneously dig deep to explore the functions and design the required functional allocation logic.

What are the risks associated with becoming over-reliant on AI? Under-reliance represents inefficiency, while over-reliance represents risk. This question raises the importance of understanding the true trustworthiness of an AI and its capability, and of conducting a continuous assessment of its behaviour to calibrate levels of trustworthiness. Using a risk lens, the answer would also lie in the frequency of being over-reliant, the magnitude of consequences and the trustworthiness of the AI. Over-reliance will lead to negative consequences. The more negative consequences we see, the more likely we will distrust an AI and remove it from service.

Can we design triggers for intervention? Should humans intervene in the functioning of the AI? If yes, when should they? A good starting point towards answering these questions is the literature on function allocation. Different forms of human–AI relationships call for different answers to these questions. If the AI is skilled in a task where it is supervising a human, it would be inefficient for the human to intervene in the AI's task. In a supervisory control task where the human supervises the AI, clearly the human is authorised and exists to intervene with the AI when needed.

E-cookies aim to encode triggers for intervention in the allocation agent. E-cookies need to be designed as a smart watchdog to ensure that the AI is performing correctly. However, this is not a simple expectation. There are fundamental challenges in the design of E-cookies that need rigorous science to address them, for example, how to assess the safety of an AI that continues to learn and evolve with new behaviours that did not exist at the design stage.

7 Conclusion

In this article, I have discussed the human–AI relationship using scientific and technological lenses. The literature of function allocation has shown that the human–AI relationship is not only about humans using AI or humans interacting with one thing called the AI, but also about different forms of micro-relationships that involve functional interactions. In a semi-autonomous car, the human interacts with the displays inside the car, the wheel when it needs to, the navigation system and so on. Similarly, in fully autonomous cars, the human will interact with the car using voice, receiving audio and visual information, and possibly haptic feedback. Each of these interactions performs different functions and engages in services offered to the human. The human trust in the car will be influenced by these different interactions.

I have argued that the Human-AI relationship should not be studied from a mere use perspective alone, which could lead to sever negative consequences. I discussed the allocation agent representing an AI that is dynamically reallocating functions between the human and the AI. The allocation agent itself is an AI and could simply cause the human to distrust or mistrust the AI in the car if the allocation agent is not skilled enough to reallocate functions effectively and efficiently. It is important to design the allocation agent, and clearly articulate the level of delegation to that agent and its authority to act in different contexts. The AI in the future could turn out to be AIs within an AI, making it difficult to trace the root causes of distrust or mistrust between the human and the machine.

I conclude that any social integration of AI into the human social system would necessitate a form of relationship on one level or another between the human and the AI in society, meaning that humans will 'always' actively participate in some decision-making loops that will influence the operations of AI. Even the most autonomous and clever AI will exist within a social system in which it needs to interact with humans and other AI systems. An AI must become socially integrated.

8 Acknowledgement

I would like to acknowledge the valuable feedback from members of my research team Essam Debie, Aya Hussein and Raul Fernandez Rojas. This research is supported by an Australian Research Council discovery project DP160102037.

9 References

Abbass H. A. (2015). *Computational red teaming: Risk analytics of big-data-to-decisions intelligent systems*. Switzerland: Springer.

Abbass, H. A. (2017). An AI professor explains: Three concerns about granting citizenship to robot Sophia. *The Conversation*. Retrieved from https://theconversation.com/an-ai-professor-explains-three-concerns-about-granting-citizenship-to-robot-sophia-86479

Abbass, H. A., Petraki, E., Merrick, K., Harvey, J., & Barlow, M. (2016). Trusted autonomy and cognitive cyber symbiosis: Open challenges. *Cognitive Computation*, *8*(3), 385–408.

Abbass, H., Tang, J., Amin, R., Ellejmi, M., & Kirby, S. (2014). The computational air traffic control brain: Computational red teaming and big data for real-time seamless brain-traffic integration. *Journal of Air Traffic Control*, *56*(2), 10–17.

Adrian, E. D., & Matthews, B. H. (1934). The Berger rhythm: Potential changes from the occipital lobes in man. *Brain*, 57(4), 355–385.

Babcock, M. L. (1960). Reorganization by adaptive automation (Doctoral dissertation), University of Illinois at Urbana-Champaign.

Berger, H. (1929). Electroencephalogram in humans. *Archiv fur Psychiatrie und nervenkrankheiten*, *87*, 527–570.

Berry, P. C. (1961). *Psychological study of decision making* (Technical Report NAVTRADEVCEN 797-1). Arlington, VA: Psychological Research Associates.

Bringsjord, S., & Schimanski, B. (2003, August). What is artificial intelligence? Psychometric AI as an answer. In Proceedings of the International Joint Conference on Artificial Intelligence, 887–893.

Controzzi, M., Cipriani, C., & Carrozza, M. C. (2014). Design of artificial hands: A review. In: Balasubramanian R., Santos V. (Ed.), *The Human Hand as an Inspiration for Robot Hand Development*. Springer Tracts in Advanced Robotics (pp. 219–246), 95. Cham, Switzerland: Springer.

Demos, J. N. (2005). Getting started with neurofeedback. WW Norton & Company.

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation*, (pp. 275-320). Oxford, England: Univer. Nebraska Press

Deutsch, M. (1973). The resolution of conflict: Constructive and destructive processes. American Behavioral Scientist, 17(2), 248-248.

Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. (2017). Microservices: Yesterday, today, and tomorrow. In Mazzara M., Meyer B. (Ed.), *Present and Ulterior Software Engineering* (pp. 195–216). Cham, Switzerland: Springer.

Dunne, P. E., Wooldridge, M., & Laurence, M. (2005). The complexity of contract negotiation. *Artificial Intelligence*, *164*(1–2), 23–46.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27.

Endsley, M. R. (2018). Automation and situation awareness. In Parasuraman, R., & Mouloua, M. (Ed.), *Automation and Human Performance* (pp. 183–202). Routledge.

Endsley, M. R. & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.

Fetzer, J. H. (1990). What is artificial intelligence? In Fetzer, J. H. (Ed.), *Artificial intelligence: Its scope and limits* (pp. 3–27). Dordrecht, The Netherlands: Springer.

Fitts, P. M., Viteles, M. S., Barr, N. L., Brimhall, D. R., Finch, G., Gardner, E., … Stevens, S. S. (1951). *Human engineering for an effective air-navigation and traffic-control system* (and appendixes 1–3). Columbus, OH: Ohio State University Research Foundation.

Flegg, G. (Ed.). (1989). *Numbers through the ages*. Palgrave, London: Macmillan International Higher Education.

Goh, S. K., Abbass H. A., Tan K. C., Al-Mamun A., Thakor N., Bezerianos A., & Li, J. (in press) Spatio-spectral representation learning for electroencephalographic gait pattern classification. *IEEE Transactions on Neural Systems & Rehabilitation Engineering*.

Hainsworth, D. W. (2001). Teleoperation user interfaces for mining robotics. *Autonomous Robots*, *11*(1), 19–28.

Hill, J. M. (1973). *Biocybernetics Project*. Cambridge, MA: Computer Corporation of America.

Inagaki, T. (2003). Adaptive automation: Design of authority for system safety. *IFAC Proceedings Volumes*, *36*(14), 13–22.

Kim, G. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin*, *68*(2), 104–120.

Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a 'team player' in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91–95.

Kovatchev, B., Cheng, P., Anderson, S. M., Pinsker, J. E., Boscari, F., Buckingham, B. A., ... Chernavvsky, D. (2017). Feasibility of long-term closed-loop control: A multicenter 6-month trial of 24/7 automated insulin delivery. *Diabetes Technology & Therapeutics*, *19*(1), 18–24.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), 4–11.

Luhmann, N. (1979). *Trust and power*. Chichester, UK: John Willey & Sons.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology, 40*(1–2), 187–195.

Randell, B. (Ed.). (2013). *The origins of digital computers: Selected papers*. Springer.

Reddy, R. (2006). Robotics and intelligent systems in support of society. *IEEE Intelligent Systems*, *21*(3), 24–31.

Rouse, W. B. (1994). Twenty years of adaptive aiding: Origins of the concept and lessons learned. In: M. Mouloua and R. Parasuraman. Eds., *Human performance in automated systems: Current research and trends*. (pp. 28-32), Hillsdale, NJ, Erlbaum.

Rouse, W. B., & Rouse, W. B. (1991). *Design for success: A human-centered approach to designing successful products and systems.* New York, NY: Wiley.

Schmorrow, D. D., & Kruse, A. A. (2002). DARPA's augmented cognition program – tomorrow's human computer interaction from vision to reality: Building cognitively aware computational systems. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*, 7–7.

Schmorrow, D., Stanney, K. M., Wilson, G., & Young, P. (2006). Augmented cognition in human-system interaction. In G. Salvendy (Ed.). *Handbook of human factors and ergonomics* (3rd ed., pp. 1364-1383). Hoboken, New Jersey: John Wiley & Sons.

Scholtz, J. (2003, January). Theory and evaluation of human robot interactions. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences,* 10-pp.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA, USA: MIT press.
Sheridan, T. B. (2012). Human supervisory control. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (4th ed., pp. 990–1015). Hoboken, New Jersey: John Wiley & Son.
Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Massachusetts Inst of Tech Cambridge Man–Machine Systems Lab.
Smallegange, J. A. P., Bastiaansen, H. J. M., Venema, A. P. & Bronkhorst, A. W. (2018) *Big data and artificial intelligence for decision making:* Dutch Position Paper, technical report STO-MP-IST-160. NATO.
Susman, R. L. (1991). Who made the Oldowan tools? Fossil evidence for tool behavior in Plio-Pleistocene hominids. *Journal of Anthropological Research, 47*(2), 129–151.
Taylor, R. H., Menciassi, A., Fichtinger, G., Fiorini, P., & Dario, P. (2016). Medical robotics and computer-integrated surgery. In R.H. Taylor (Ed.), *Springer handbook of robotics* (pp. 1657–1684). Cham, Switzerland: Springer.
Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.
Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering, 2*(1), 157–180.