frontiers | Frontiers in Physics

TYPE Review
PUBLISHED 07 November 2022
DOI 10.3389/fphy.2022.951296

OPEN ACCESS

EDITED BY
William Frere Lawless,
Paine College, United States

REVIEWED BY
Michael Wollowski,
Rose-Hulman Institute of Technology, United States
Yohei Katano,
Meiji University, Japan

*CORRESPONDENCE
Mito Akiyoshi,
mito.akiyoshi@gmail.com

SPECIALTY SECTION
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

RECEIVED 23 May 2022
ACCEPTED 24 October 2022
PUBLISHED 07 November 2022

CITATION
Akiyoshi M (2022), Trust in things: A review of social science perspectives on autonomous human-machine-team systems and systemic interdependence. *Front. Phys.* 10:951296. doi: 10.3389/fphy.2022.951296

Trust in things: A review of social science perspectives on autonomous human-machine-team systems and systemic interdependence

Mito Akiyoshi*

Department of Sociology, Senshu University, Kawasaki, Japan

For Autonomous Human Machine Teams and Systems (A-HMT-S) to function in a real-world setting, trust has to be established and verified in both human and non-human actors. But the nature of "trust" itself, as established by long-evolving social interaction among humans and as encoded by humans in the emergent behavior of machines, is not self-evident and should not be assumed *a priori*. The social sciences, broadly defined, can provide guidance in this regard, pointing to the situational, context-driven, and sometimes other-than-rational grounds that give rise to trustability, trustworthiness, and trust. This paper introduces social scientific perspectives that illuminate the nature of trust that A-HMT-S must produce as they take root in society. It does so by integrating key theoretical perspectives: the ecological theory of actors and their tasks, theory on the introduction of social problems into the civic sphere, and the material political economy framework developed in the sociological study of markets.

KEYWORDS

machine, algorithm, artificial intelligence, interdependence, sociology, trust

1 Introduction

In this paper, Autonomous Human Machine Teams and Systems (A-HMT-S) are defined as teams that include humans and increasingly intelligent and autonomous machines working together [1, 2]. Intelligent machines are defined as machines or algorithms that think by scanning data for patterns, make inferences, and learn by testing inferences [3]. Advances in deep learning in the 21st century bring this emerging phenomenon closer to reality [1, 2, 4], though the idea of thinking machines was explored decades ago by Turing, Shannon, Weiner, Simon, and others, and was foreshadowed to an extent by Babbage's Difference Engines and Analytical Engine a century before that [5].

A key advance in the conceptualization of A-HMT-S is that intelligent machines are intended to operate as full-fledged team members collaborating with humans [4, 6]. Not only do they assist human decision-making and automate information processing, they also make decisions on their own and instruct human workers and other machines [7].

For example, an artificially intelligent co-worker named Charlie has been developed by Cummings et al. [4]. Charlie is designed to perform typical white-collar tasks: she gives interviews, takes part in brain-storming sessions, and collaborates in writing papers. But exhibiting recognizable and anthropomorphized agency or human-like identity, as Charlie does, is not central to the definition of A-HMT-S. They may not have the attractive features of *Blade Runner*'s replicants, but they have the kind of intelligence that would pass the Turing Test in the specific tasks to which they are assigned. They will also someday pass the "toilet test"—the ability to run unsupervised while humans address their bodily needs [8]. In short, the defining feature of intelligent machines that constitute A-HMT-S is that they can model human recognition, learning, and reasoning [3].

As our machine helpers become increasingly autonomous and intelligent, it leads to increasing interdependence between human and non-human actors. Although that evolution is far from complete and may never end, quasi-A-HMT-S with semi-autonomous machines are now commonplace. They diagnose and treat diseases [9], drive vehicles [1, 10], fly airplanes [11, 12], educate students, conduct research [4], trade stocks and derivatives [8], market products and services [13], fight wars [2], all with mixed results.

Algorithms lie at the core of these capabilities. In addition to their sheer ubiquity, their complexity and opacity and their anticipated consequences for humanity have motivated interdisciplinary research on societal effects of A-HMT-S [14, 15]. This paper is part of that interdisciplinary effort, addressing a crucial aspect of the implications of the integration of A-HMT-S into society: trust.

In traditional organizations, trust among workers is essential in achieving quality performance [16]. The increasing interdependence between humans and intelligent machines poses a series of trust-related questions: as machines become more autonomous, what are the causes and consequences of trust-building in A-HMT-S? What does it mean to trust non-human actors in a system? This paper uses a social-scientific toolkit to address these questions. In doing so, it might help to quickly look backward for a moment and consider the issue as it was faced by users of the earliest known human tools: handaxes. The user of a handaxe had to "trust" that its shape and material would be adequate to the task, which usually involved cutting into some kind of organic matter. Since the user probably also made the tool, she or he had an inbuilt basis for trusting it, including to trust that it wouldn't suddenly assume agentive power of its own and diverge from the user's goal, notwithstanding any animistic beliefs that might have been in play. The only other entities with "agency" in this scenario would have been other proto-humans, and the distribution of trust across the group would be established by longstanding social norms and rules. In short, the issue of trust was severable from other considerations, and its resolution was an intra-human one.

The history of technology since then has seen that simple allocation of trust be thoroughly complicated by the folding of more and more human capability into the tools themselves—at first physical and then mental [17]. A late medieval cannoneer had to trust the cannon wouldn't blow up in his face, but the location of that trustworthiness still resided in the cannon-maker. A Jacquard loom weaver, on the other hand, didn't have to place her trust in the card-maker because the output of the loom would reveal if the card-punching was accurate. A paddle-wheel steamboat passenger had to trust the boat and its crew, but might have known nothing about the mechanical steam engine governor that could be trusted (usually) to keep the engine speed steady. Today's automobile driver may only partially grasp the extent to which their survival depends on the trustability of dozens of microchips installed in the vehicle by factory workers who were overseeing relatively simple robots, which in turn had to be trusted to work right, with that chain of trust extending all the way back to the machines that designed the machines that designed them. Trust, once a human prerogative, is now diffused across multiple overlapping systems of systems. A-HMT-S is the inheritor of this long process.

But what is trust, and what makes an entity trustworthy? This paper accepts a widely agreed-upon definition of trust as the willingness of a trusting entity (the trustor) to be vulnerable to a trusted entity (the trustee) with respect to a pertinent domain, a trust object, against a backdrop of risk and uncertainty. Trust is therefore not a static thing but a constantly changeable relationship between actors, based on the assessment of each other's behavior in the relationship. One or both parties have just enough evidence to believe that the relationship will work out the way each of them expects it to [18–20]. Though fragile, it is an absolute, foundational basis of society. That is why Dante in his *Inferno* reserved the lowest circle of hell for people who have betrayed other people's trust. Trustworthiness, meanwhile, is a roughly quantifiable set of properties that the trustee in a relationship displays to the trustor to signal their intentions and probable behavior.

Each dimension of trust—trustor, trustee, and trust object—is expressed across a spectrum of generality ranging from the most particular to the most highly generalized [18]. For example, one terrible visit to a physician may imply the withdrawal of trust in that particular doctor, in the category of medical professional she or he represents (e.g., cardiology), or in the entire community of medical experts. Whether a particular visit results in the demise of trust at any level of generality depends on other pertinent variables.

From that starting point, this review will provide a synthesis of key social scientific thinking relevant to the question of trust within and between human and non-human actors. The next section reviews social scientific literature on interpersonal trust, which is compared with human-machine trust in the section after that. Empirical and experimental studies have shown that multiple factors including algorithmic transparency and

machine error rates affect the level of trustworthiness that humans ascribe to intelligent machines [21]. But trust in A-HMT-S is not fully reducible to design issues; we will see that the broader context of interactions between A-HMT-S and other spheres of society is also relevant. In order to examine inter-system trust, the fourth section draws on the urban ecology tradition in sociology, as well as on research on the construction of social problems and the sociology of technology. But rather than introducing concepts in the abstract, it discusses specific incidents that involve precursors of A-HMT-S. By way of conclusion, this paper argues that the issue of trust in A-HMT-S is a specific case of the broader issue of trust in abstract systems and that as such, trust-building spans multiple social ecosystems and is supported or undermined by interactions among them.

2 Industrialization and the transmutation of trust

2.1 Interpersonal trust

Interpersonal trust is a linchpin of society. As discussed in Section 1, trust processes can be analyzed in terms of the trustor, the trustee, the trust object of varying generality degrees. Small-scale societies are characterized by particularized trust because interactions tend to be embedded in a local context [17]. Societies that are more complexly organized require coordination among actors we may not personally know; in such societies, reliance on general trust has become widespread and is essential to their continued existence [22]. In either case, trust depends on complex mutual understandings that defy easy definition [23]. This tacit and yet robust trust in others to do what a mesh of overt and latent rules dictates, and which makes the social order possible, is one major focus of ethnomethodology, the sociological and anthropological study of the rules by which people organize their everyday lives [24].

Interpersonal trust, in this perspective, operates on a provisional basis, and involves a sort of pattern-matching exercise. Confirming every datum imaginable and eliminating all alternate interpretative possibilities are neither possible nor called for unless the veracity of a person's explicit or tacit claim is called into question. A just-good-enough assessment of the situation suffices [24]. Thus, if someone who "looks like a college professor" enters a college classroom and approaches the podium, students assume that person is the course instructor and rarely ask for official proof of his or her identity. Additional elements of legitimation may appear in the form of references to the shared institutional structure that encompasses both the professor and the students—the topic of the course, the academic calendar, the grading system. As long as the behavior matches the observer's expectations in that setting, provisional trust will be satisfied.

We all do this a hundred times a day without even thinking about it. Social interaction is made possible by everyone's taking everyone else's claims at face value unless some contradictory evidence emerges that requires vetting [23]. The taken-for-granted nature of social life constitutes a cognitive and emotional common ground that is prior even to shared values and norms—things that are thought of as "culture" in the social scientific sense. Trust evolves over time in organizations through interactions that involves people's values, attitudes, and emotions [16].

Because interpersonal trustworthiness is not fully or even primarily grounded in the procedure of fact checking, societies vary widely in terms of the level of confidence people have about one another [25]. This is verifiable by looking at situations where it is lacking. For example, the mafia-type organized crime syndicates in southern Italy came into being as enforcers of contracts in a low-trust environment [26, 27]. Farmers who could not trust their counterparties in selling or buying produce and livestock had to turn to proto-mafiosi to guarantee transactions with threats of violence. Similarly, neighborhoods with high crime rates must invest heavily in security, and endure stressful anxiety, whereas individuals in low-crime areas can insouciantly leave their doors unlocked when they go out to run errands. The erosion of trust makes lives difficult. Until destroyed, the operation of trust tends to remain invisible, and yet trust is a public good from which other advantages such as cooperation, tolerance, functioning democracy, and market efficiency come about [16, 28].

2.2 Trust in machines and abstract systems

Industrialization extended the scope of trust relationships to include abstract systems [29]. Individuals and organizations in highly industrialized societies must learn to trust knowledge systems and technologies they do not fully grasp. Again, perfect grounding is precluded and faith is an integral dimension underlying trust. People board trains not knowing how the public transportation system is organized and operated, and they receive mRNA vaccines to protect themselves against viral infections without a detailed understanding of the immune system or vaccine manufacturing. Workers also learn, through trial and error, to trust machines they operate to mass produce goods and services. The threat of deskilling might be seen as a potential source of the erosion of trust in cases of automation, but Zuboff finds that workers adopt and adapt through explorative use of new technologies and achieve reskilling by becoming their adept and creative users [30].

In our capacity as consumers, too, we have entered a world where we buy things produced by distant others. The rise of advertising and branding is associated with this shift towards mass production, distribution, and consumption which Beniger has called "the control revolution" [17]. Advertising and

branding are important where interpersonal trust cannot guarantee the quality of goods produced by large-scale organizations and sold anonymously. As Max Weber's celebrated analysis has shown, bureaucracy arises to enable the operation of such organizations by releasing trust from the domain of interpersonal relationships and the immediacy of face-to-face interaction, replacing it with formally defined rules and procedures and a hierarchy of offices [31].

3 Difficulties of building trust in A-HMT-S

Although trust in A-HMT-S has unique aspects, in principle the questions it raises are predictable extensions of the centuries-long process that preceded it [29]. Prior to the development of A-HMT-S, there were systems consisting of human operators and non-autonomous and non-intelligent machines and tools: vehicles, missile systems, nuclear power plants, and so on [1, 17]. I call these complex but non-intelligent tools "mundane systems" in contrast to A-HMT-S.

Technology scholars Hengstler, Enkel, and Duelli argue that trust in automated systems has two aspects: trust in the automation technology itself and trust in organizations that develop it, use it, or in which it is embedded [32]. However, in the case of trust in A-HMT-S, it is neither analytically tractable nor appropriate to separate the technology from its organizations and institutions. The literature on the sociology of technology has demonstrated the futility of treating a technology's capabilities without reference to its users and its context of use. According to the constructionist perspective of technology, there is no such thing as technology per se [33, 34]. The emergence of A-HMT-S reasserts that point with renewed exigency: in A-HMT-S, the technology implements, enacts, and embodies organizations' purposes and goals. Technology is the organization in a literal sense, and *vice versa*.

Shestakofsky conducted participant-observation research at a software firm and found that two types of labor were performed to create dynamic collaboration between humans and autonomous algorithms [35]. Computational labor addresses the issue of machine lag, problems posed by limitations of technologies. Human teams engage in repetitive information-processing tasks in order to fix gaps in software infrastructure. At the same time, emotional labor by human workers deals with human lag, clients' reluctance to use algorithms, and mediates the relationship between software systems and the latter. These findings suggest that trust among A-HMT-S actors is constructed in the course of collectively defining tasks and negotiating boundaries [35]. Jarrahi argues that human-AI symbiosis in organizational decision-making is possible when AI supplements human cognition and humans bring a holistic and intuitive approach in dealing with uncertainty [36].

A theoretical framework that addresses the issue of trust in A-HMT-S may be developed by treating the amalgam of non-humans and humans as-they-are. Studies have shown that human-to-machine trust is affected by various factors: the extent to which the machine exhibits human-like appearance, cognitive biases in general, automation-specific complacency and bias [37], algorithmic error rates, epistemic opacity, and the type of tasks [38]. Trustworthiness can be ascribed to intelligent machines and form a basis of productive collaboration in A-HMT-S, but the presence of biases and complacency means that humans can over-trust or under-trust intelligent algorithms and their decisions.

The problem with A-HMT-S is that it often involves "black box algorithms," epistemically opaque to human observers because they keep self-improving by testing and learning [9]. Opacity raises concerns among developers, users, and the general public. Lee and See, observing that trust is essential in the adoption of automation systems, recommends such measures as the disclosure of intermediate results of the algorithms to the operators and the simplification of algorithms [20]. Similarly, Burrell supports greater regulations, algorithmic transparency, and education of the public [9, 39]. The Defense Advanced Research Projects Agency (DARPA) attempted to address the opacity issue by developing "explainable artificial intelligence" [40]. Whether systems that "look" human, or visibly inserting actual humans into the decision loop, have any effect on trust and affinity is also investigated [41, 42]. It is important, though, to recall that the issue of trust in "black-box algorithms" is only among the latest developments in the history of trusting increasingly distant others and longer chains of factors.

Durán and Jongsma argue, using medical AI as a case study, that trust in black-box algorithms can be established by the principle of computational reliabilism (CR) [9]. Striving for algorithmic transparency, they claim, is a losing strategy because it defeats the purpose of deploying algorithms in the first place. "Transparency will not provide solutions to opacity, and therefore having more transparent algorithms is not a guarantee for better explanations, predictions and overall justification of our trust in the result of an algorithm" [9, p.331]. They suggest employing a version of the heuristic devices we use to assess the trustworthiness of our social interlocutors. In any given setting, CR assesses the trustworthiness of AI not by using interpretive parameters to check the system's inner state at points 1 through n, but by making multiple empirical inferences that turn out to be "good enough": A comparison with known solutions (verification), comparison with experimental data (validation), robustness analysis, a history of successful or unsuccessful implementation, and expert knowledge. An analogy with human interaction is to judge people by their behavior and set aside speculation about the mental processes that led to that behavior. Epistemological opacity does not have to be removed as long as CR can be established [9]. This enables

users to take advantage of sophisticated black box analysis while solving the dilemma of being dependent on it without comprehending its workings.

This is particularly important for medical AI, but is applicable to other domains and to the question of building trust in non-AI abstract systems. It is similar to the satisficing that we saw in the college professor story earlier. Limited as we all are by bounded rationality [3, 43, 44], humans and organizations have to abandon the ideal of perfect explainability and treat the state of trust as provisional and dynamic. Yet for this very reason provisional trust is a fragile construct that can collapse if challenged by outsiders. And that is likely to happen at the border between A-HMT-S and other communities across the broader society with which it interacts. At that interface, CR may not help. To address the fact that heterogeneous actors scattered across heterogenous fields also will be asking themselves questions about the trustworthiness of A-HMT-S, and about the impact of A-HMT-S on their own interests, the next section turns to the ecological perspective originated in urban sociology.

4 A-HMT-S as an ecosystem

Establishing trust in A-HMT-S increasingly entails ethical as well as legal challenges, including transparency, algorithmic fairness, safety, security, and privacy. Challenges in jurisprudence emerge when non-human actors assume human-like characteristics. Scientific as well as practitioner knowledge systems engage in articulating goals and means in trust promotion and production [45]. Opening up black-box algorithms is often presented as a key to this undertaking. But as we have seen, perfect algorithmic transparency is not always feasible or effective. To identify and better understand trust goals relevant to A-HMT-S, an urban ecology perspective is useful. Urban ecology, a sociological perspective developed by scholars at the University of Chicago in the 1920s allows us to grasp the dynamic and emergent nature of the trustor and the trustee in interaction because it incorporates heterogenous actors and can incorporate A-HMT-S as a focus of trust processes. Borrowing its key metaphor and related concepts, such as territorial competition and inter-group cooperation, from biology, it sought to account for the ways different populations distributed themselves across the space of the city and used its resources. In that tradition, authors sometimes use the word "ecology" to describe what we conventionally understand by the term "ecosystem" [8, 46]. To avoid confusion, this paper will use that more conventional term. An ecosystem is an autonomous domain of actors, their tasks, and the relationship between actors and tasks [46]. It also includes the resources they obtain from the environment, and the other ways they interact with their surroundings. Territorial shifts of populations are seen in terms of invasion and ecological succession or the replacement of one group by another. For example, residential patterns of immigrants to major cities in the United States at the turn of the 20th century were determined by their place of work—often in the central business district—, as well as by their material means, and their social distance from native populations. Neighborhoods that had seen the arrival of immigrants experienced an exodus of middle-class families; the new groups further affected the types of businesses and services in these transitioning neighborhoods. The distribution of populations and differentiation of space are subjected to the process of interaction among diverse groups.

At this level of analysis, we can think of whole ecosystems as units of interaction. A-HMT-S researchers, developers, and popularizers constitute one such ecosystem. For people outside it to trust "what the machines are doing," they have to trust or at least tolerate the ecosystem as a whole, including the motivations and behavior of the humans, the type and amount of environmental resources it uses and the way it uses them. Outsiders have to satisfy themselves that none of this poses a threat to their individual and collective livelihood or to how they understand the world and act in it. And they have to figure out how to minimize friction at the interface between their own ecosystem and that of the newcomer. As was mentioned earlier, achieving and keeping a state of trust will bring both cognitive and emotional dimensions into play, and the benchmark will tend to be: How well does this new ecosystem play by the taken-for-granted rules of everyday life [24]?

In the case of medical A-HMT-S, for instance, in order to take root in day-to-day medical practice it has to build trust relationships with patients, regulators, healthcare providers, insurance providers, and the general public. Computational reliabilism may be a necessary but not sufficient condition for that, as each party may judge the situation by different criteria. Physicians may be most concerned with diagnostic accuracy while insurance providers may worry over the cost-benefit issues and hospital technicians may care about fitting new practices into existing routines. If we recall that trust is a relation of varying generality as discussed in Section 1, then highly particularized trust in a trust object does not entail trust in a category or ecosystem of which that trust object is an instantiation. A particularized trust object is in fact a construct of multiple ecosystems. Society-wide trust in A-HMT-S is thus a constant balancing act. And as we will see in a later section of this paper, it can be lost when a failure occurs and the system as a whole does not engage in trust-repairing behavior addressed collectively to people living and working in other ecosystems.

Mackenzie, drawing on Abbott, used the ecosystemic perspective in a study of the rise of High-Frequency Trading and its relation to existing trading and regulatory systems [8]. His research reveals the ripple effect of technological decisions as they impinge on the interests of other domains. HFT is a type of A-HMT-S made possible by machines that can analyze opportunities and execute orders at a speed that surpasses that of human-only teams. Because of this advantage, HFT

firms quickly became major players in their respective markets. In the process, they generated enormous profits by engaging in legal but arguably unscrupulous trading activities, made possible only by the high-speed of their systems. Then, in a move apparently unrelated to what the HFTs were up to, the New York Stock Exchange decided to install a new communication antenna on the roof of its data center. Available to any member who paid the requisite hefty fee, the antenna would provide a half-microsecond improvement in transaction time by eliminating 260 m of fiber optic cable from the transmission line. This was exactly the sort of time difference the HFTs had been exploiting through their proprietary technology, and now their advantage was threatened.

As a prelude to explaining what ensued, MacKenzie revisits an insurrection that took place in the English community of St. Albans in the late 14th century [8]. As part of a general wave of uprisings against feudalism, townspeople invaded the local Benedictine monastery and, after freeing people held in the monastery's prison, entered the abbot's parlor, methodically smashed its stone-paved floor, and carried pieces of it away with them. This seemingly random act was in fact retaliation for a previous abbot having confiscated the townspeople's millstones 50 years earlier and used the confiscated stones to pave the parlor floor. The motive for that had been to achieve a monastic monopoly over grain-milling and extract the consequent fees. Townsfolk never forgot this, which exemplifies a key point MacKenzie wants to emphasize: even seemingly minor changes in available technology are not neutral but are usually bound up in power relations with long-lasting effects.

Back in the 21st century New York, the new antenna plan had similar consequences that drew in multiple institutional spheres—which MacKenzie refers to as "ecologies." Eventually, the Securities and Exchange Commission, a local zoning board, residents of the town where the data center is located, the Stock Exchange itself, and others found themselves in conflict over something which had seemed like a simple technology decision: eliminating 260 m of fiber. The eventual solution once again exemplifies the ways in which a material consideration can be waylaid by issues of power: as of 2020, it had been decided to reinsert the half-microsecond delay by adding a coil of cable to the transmission line, thereby returning everything to the status quo ante.

Mackenzie's point is generalizable. Just as biological populations compete for habitat and resources, different social actors behaving collectively will interact to create an observed distribution of functions (tasks that need to be executed for the maintenance of order) and habitats within and between ecoystems. Interactions will define actors and the nature of their tasks; what gets done, and who does it, are not rigidly defined by pre-existing functions [46]. Instead, turf battles for resources and legitimacy dynamically shape the things actors do and don't do, in a manner that social scientists call "co-constitutive" and that other disciplines might term "emergent." Squabbling over a length of fiber optic cable, and expropriating a paving stone, can be inexplicable outside of a specific social, political and economic context that makes them highly meaningful.

The rapid growth of A-HMT-S capabilities and governmental attempts to control that process is another part of this story of ecosystems squaring off against one another. Whether unfettered development is encouraged or restrained is a function of interactions among the affected ecosystems. Lethal autonomous weapons systems (LAWS) provide a good example [2]. They will proliferate in a society if other ecosystems that interact with it invest in and legitimize their development, but will be suppressed in any society where the state reins in the military deployment of A-HMT-S.

The above examples show that when A-HMT-S is deployed it can trigger social effects across multiple domains. In the labor market, it can result in job creation, job elimination, or both. In the political domain, it can produce a crisis among regulators and legislators. Pfeffer addresses such broader implications in a study of the impact of AI on the economy and workers' well-being [47]. He points out that the introduction of A-HMT-S can have detrimental effects on workers by eliminating jobs and forcing some workers to switch occupational categories, many of whom already experience stagnant wages and job precarity. Low fertility, government budget deficits, and runaway debt in many highly industrialized societies mean that public policy interventions to attenuate the negative labor market impacts of A-HMT-S are unlikely. A-HMT-S can be used to promote human well-being, but Pfeffer observes that they are as likely to be used in ways that exacerbate economic inequities [47]. If workers come to regard A-HMT-S as a tool to make themselves redundant, computational reliabilism will probably not help them trust it.

The expanding use of A-HMT-S will also force revisions of school curricula, similar to the way basic computer skills became a key subject in the final decades of the 20th century [48]. One can envisage a future in which students are required to learn how to work with A-HMT-S to optimize learning. The ecosystemic perspective helps us understand the complex nature of systems interacting with their environments; it enables us to see that what seems external to systems themselves are in fact constitutive of their functions. Adjacent ecosystems regulate, offer incentives and resources, call for accountability, and do many other things that can influence the success of A-HMT-S.

In terms of its effects on human activity, A-HMT-S is more than the automation or translation of tasks formerly performed by humans. It leads to the emergence of new tasks to address the challenges that it and other ecosystems present to each other as they each seek to thrive in the world they must share. In the course of building explainable systems, A-HMT-S must also explain itself to any audience whose activities could be upended by it. At first glance, it may have seemed strange that Pfeffer's paper on the effects of AI has data on fertility,

national deficits and debts, but the ecosystemic perspective motivates such a focus on a nexus of multiple spheres [47].

5 How technological systems can breach trust

Prior to the development of A-HMT-S, there were many systems made up of human operators and non-autonomous and non-intelligent machines and tools: vehicles, missile systems, nuclear power plants, and so on. I referred earlier to these non-intelligent tools as "mundane systems" in contrast to A-HMT-S. Mundane systems have a track record of breaching the trust of their users and the general public. The way they fail illuminates the kind of trust issues that A-HMT-S may face going forward.

5.1 Mundane system trust erosion: Three brief examples

Drunk driving: Car accidents caused by drunk drivers, and the public discourse surrounding them, remind us that the accepted narrative of interdependence between driver, car, and environment is only one of several potential ways to constellate the relevant elements. Typically, when an accident happens the drunk driver is designated as the "cause" and becomes the target of moral opprobrium. Alternate reasonings are possible but rarely accepted in what Gusfield calls the public drama of social problems [49]. The lack of public transportation to venues that serve alcohol, or the mingling of cars and pedestrians on the same thoroughfares, could be conducive to accidents caused by drunk driving, and yet poor urban planning is rarely singled out as a cause. Car manufacturers are not held accountable for building vehicles that can kill regardless of what mental state the operator is in. The underlying assumption regarding the interdependence of the driver, the car, and the streets is that the driver should be a morally upstanding individual who exercises prudence and is capable of controlling their own behavior. The presence of accidents caused by sober but incompetent drivers indicates that the association between behavior and morality involves the choice of a certain perspective.

Titan II missile explosion: In 1980, a Titan II intercontinental ballistic missile at a missile complex in Damascus, Arkansas was damaged when a worker accidentally dropped a wrench socket down its silo during a routine check of the oxidizer tank pressure, which caused a fuel leak [50]. The fuel exploded the following day, resulting in one death and multiple injuries. The interdependence of humans and non-intelligent machines can go awry without moral failure by the humans. The coexistence of the worker, the socket, and the vulnerable tank surface led to the explosion.

Fukushima Daiichi Nuclear Power Plant failure: After the East Japan Earthquake of 2011, the resulting tsunami hit the Fukushima Daiichi Nuclear Power Plant and its reactor cooling system failed. This led to reactor meltdowns, explosion and the atmospheric release of radioactive material [51]. A nuclear plant is an example of a mundane system. Even though the plant uses multiple machines and robots, they are not autonomous or intelligent. In the case of the Fukushima Daiichi Nuclear Powerplant, it turned out that TEPCO, the plant operator, and other related organizations had underestimated the risk of losing reactor cooling after a tsunami. Some seismologists familiar with the region's earthquake and tsunami history had warned that a cooling system failure due to major tsunami was possible, but those warnings were not heeded [52]. The interdependence between humans and the plant was disrupted not by a gap intrinsic in their relationships—both humans and the plant were executing tasks assigned to them—but by TEPCO management's decision years earlier to ignore evidence of a serious environmental risk.

As these cases illustrate, the interdependence of elements in mundane systems can be eroded by various factors. The misplacement of trust may only become evident ex post. Drunk drivers should not be trusted to drive safely and yet there is currently no scalable solution to prevent them from getting behind the wheel. The missile fuel tank was not designed to withstand the damage caused by a falling wrench socket, and it was never anticipated that a worker might drop a socket inside the silo. The Fukushima Daiichi Power Plant was supposed to have been built on safe ground and the risk of earthquake and tsunami was believed to be manageable, because the scientists who had warned of potential damage to the cooling system were considered an untrusted minority.

Being systems comprised of human and non-human actors, and operating among other groups and systems with their own idiosyncrasies, A-HMT-S could fail in the same ways mundane systems do: lack of fail-safe mechanisms, human error, poor coordination between actors. However, they can fail in ways unique to them because they have two types of intelligence: human intelligence and machine intelligence. Some further examples will illustrate this.

5.2 Two cases of failure in systems that are "A-HMT-S-adjacent"

Boeing 737 Max: Two crashes of this Boeing model were caused by some pilots' inability to interact correctly with software that had been implemented to compensate for certain stall conditions [11, 12]. Optimistically named Maneuvering Characteristics Augmentation System (MCAS), the software conflicted with human pilots' judgement and behavioral habits acquired over years of flying previous 737s. A 737 Max without MCAS tends to nose upward in flight because of its large engines

placed high on the wing. A nose-up condition can trigger a stall, which is a bad thing for an aircraft. MCAS identifies some conditions under which it automatically forces the nose downward. In the case of the two accidents, pilots who didn't know why the plane was suddenly dipping its nose reacted incorrectly and set in motion a sequence of events that led to tragedy.

But why place the engines so high? Because more efficient engines have larger diameter than less efficient ones, and to prevent them from scraping against the ground, they had to be positioned higher on the wing than the engines on earlier 737s. This higher placement compensates for the fact that the plane's landing gear struts are short, which was a design decision made in the 1960s to make the 737 cargo bay accessible at smaller airports that lacked a full complement of loading equipment, and that design factor was never changed through many decades. A long chain of design and performance decisions, and several hundred deaths, resulted arguably from that single criterion. This also means that redesigning the wing and engine was not even possible without many other changes that would turn it into a completely new plane, requiring a lengthy and costly certification process with multiple regulatory agencies involved. Once Boeing decided to "re-engine" the 737, a software fix was the only option to compensate for the awkward aerodynamics of the high-mounted engines. Boeing vigorously lobbied with regulators to allow the design changes without fully sharing details with airline companies or pilots [11]. Pilots were not informed about the existence, much less the operation, of MCAS; in the case of the two fallen planes they had not received simulator training to work with the software.

Boeing 737 Max can be regarded as a precursor to full-fledged A-HMT-S. Humans are on the loop rather than in the loop [2]. When they are not given authority to intervene when software made a faulty move, or when they aren't sure how to react to a machine decision, the entire system fails catastrophically.

ShotSpotter: ShotSpotter uses specially designed microphones, AI, and human analysts to detect and geolocate gunshots. It claims to offer precision policing solutions to detect crimes and protect lives. In May 2020, based on evidence from this gunfire detection system, a Chicago man named Michael Williams was accused of shooting a neighbor. Forensic reports prepared by ShotSpotter employees established his culpability. After he had been in jail for nearly a year, a judge decided the evidence against him was too weak and the case was dismissed. Williams claims he was giving a ride to the victim when that person was shot by someone else [53].

As is the case with human interactions, human-machine systems must earn the trust of those with whom they interact. With the ShotSpotter case and the 737 Max disasters, these systems that are on the road to A-HMT-S may not deserve anything more than a skeptical and provisional assessment of trustworthiness. Trust in mundane systems and A-HMT-S are both examples of trust in abstract systems, which is always potentially fraught with suspicion and competing claims [29]. What is distinct about trust in A-HMT-S granted by outside actors such as the media and the political system is that it involves trust in decisions made by autonomous and intelligent machines [1, 2, 4, 7, 39]. When high-stakes decisions such as making a criminal accusation or flying an airplane are made by A-HMT-S and then turn out to be wrong, trust will naturally erode.

But A-HMT-S are not solely responsible for their ability to achieve societal trust. Other ecosystems can enhance or suppress the likelihood of it. For example, Muehlematter and Vokinger recommend that one way to improve public trust in artificial-intelligence and machine-learning-based medical devices is to increase transparancy regarding their regulation and approval. Currently, there is an unexplained gap in the timing of approval of devices commonly approved in the United State and Europe [54].

A breach in trust could also set off what Alexander called the "societalization" of A-HMT-S [55]. Societalization happens when long-enduring problems cease to be internal to a given ecosystem (in the usage we employed earlier) and are redefined as a general crisis in the public sphere. Media play the role of agenda-setter with increased and detailed coverage [55]. Investigative reporting of dramatic cases cracks them open for public discourse and denunciation. The societalization process may trigger regulatory intervention, but that will depend on whether politicians perceive that what is at stake is aligned with their own interests: another example of different ecosystems interacting at the boundary of their respective domains [46].

The 737 Max disasters and the erroneous prosecution with ShotSpotter data foreshadow what the societalization of A-HMT-S might look like. General public trust in A-HMT-S will have to be actively produced and continuously maintained if A-HMT-S is to achieve the hoped-for synergy of humans and autonomous machines. The current backlash against documented instances of biased algorithms shows the consequences of failing to secure such trust [39, 56–58]. In 2020, a computer algorithm was used to determine grades for the General Certificate of Secondary Education and A-level qualification in the United Kingdom when exams were cancelled due to the COVID-19 pandemic. The algorithm was found to disproportionately and systematically suppress the grades of students from disadvantaged backgrounds because it used the historical grade distribution at the school level to weight the grades of individual students [59]. Faced with a nationwide controversy, the algorithmically-generated grades were eventually replaced with alternative grades that integrated teachers' assessments. The emergent A-HMT-S deservedly failed to earn the trust of the public.

This section has focused on challenges involved in building trust in A-HMT-S, using cases that revealed design or deployment gaps. Of course, there are also cases in which human and non-human actors successfully achieve

fully collaborative participation. In some such cases, non-human actors acquire their own agency equivalent to that of human actors and cease to be a mere assistant to the human actors [2, 4].

6 Conclusion

This paper reviewed the social scientific literature that illuminates our understanding of issues regarding trust in A-HMT-S. In research on AI and trust, establishing trust is often presented as a matter of algorithmic transparency above all [39]. Since A-HMT-S can inadvertently incorporate existing forms of inequality and discrimination, improving algorithmic transparency is certainly a key challenge. At the same time, the present review offers a broader context. The taken-for-granted nature of interpersonal trust among humans suggests some of the ground that human-machine systems will have to cover in order to display trustworthiness, and to achieve and maintain relationships of trust [8, 23, 24]. Anthropomorphizing interfaces and developing explainable AI are attempts to achieve trust within the ecosystem of A-HMT-S. But those things alone will probably not be enough to curtail skepticism on the part of people outside that ecosystem. Skepticism is not a luddite reaction. Rather, it is a predictable caution about the effects that A-HMT-S can have on well-being of those whose lives and livelihoods may be touched by them [47, 59]. A-HMT-S researchers and developers' engagement with the labor market, academia, mass media and other domains will contribute importantly to the goal of securing trust about technologies that are not fully explicable and yet lead to highly consequential outcomes.

Author contributions

MA is solely responsible for the entire contents of the article.

Acknowledgments

I wish to thank Gerald Lombardi and William Lawless, and two independent reviewers for their insights and helpful comments.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author declares a past co-authorship with the handling editor WL.

The handling editor declared a past co-authorship with one of the authors MA.

Publisher's note

References

1. Lawless WF. Toward a physics of interdependence for autonomous human-machine systems: The case of the uber fatal accident, 2018. *Front Phys* (2022) 10: 879171. doi:10.3389/fphy.2022.879171

2. Lawless WF, Mittu R, Sofge DA, Shortell T, McDermott TA. Introduction to "systems engineering and artificial intelligence" and the chapters. In: WF Lawless, R Mittu, DA Sofge, T Shortell, TA McDermott, editors. *Systems engineering and artificial intelligence [internet]*. Cham: Springer International Publishing (2021).

3. Frantz R. Herbert Simon: Artificial intelligence as a framework for understanding intuition. *J Econ Psychol* (2003) 24(2):265–77. doi:10.1016/s0167-4870(02)00207-6

4. Cummings P, Schurr N, Naber A, Charlie SD. Recognizing artificial intelligence: The key to unlocking human AI teams. In: WF Lawless, R Mittu, DA Sofge, T Shortell, TA McDermott, editors. *Systems engineering and artificial intelligence [internet]*. Cham: Springer International Publishing (2021).

5. Gleick J. *The information: A history, a theory, a flood*. New York: Vintage Books (2012). p. 526.

6. Jiang W, Fischer JE, Greenhalgh C, Ramchurn SD, Wu F, Jennings NR. Social implications of agent-based planning support for human teams. International Conference on Collaboration Technologies and Systems (2014). p. 310.

7. Lee MK. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc* (2018) 5(1): 205395171875668. doi:10.1177/2053951718756684

8. MacKenzie D. *Trading at the speed of light: How ultrafast algorithms are transforming financial markets*. Princeton, NJ: Princeton University Press (2021). p. 290.

9. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* (2021) 47(5): 329. doi:10.1136/medethics-2020-106820

10. Panagiotopoulos I, Dimitrakopoulos G. An empirical investigation on consumers' intentions towards autonomous driving. *Transportation Res C: Emerging Tech* (2018) 95:773–84. doi:10.1016/j.trc.2018.08.013

11. Robison P. *Flying blind: The 737 MAX tragedy and the fall of boeing*. New York: Doubleday (2021). p. 336.

12. Mongan J, Kohli M. Artificial intelligence and human life: Five lessons for radiology from the 737 MAX disasters. *Radiol Artif Intelligence* (2020) 2(2): e190111. doi:10.1148/ryai.2020190111

13. Ameen N, Tarhini A, Reppel A, Anand A. Customer experiences in the age of artificial intelligence. *Comput Hum Behav* (2021) 114:106548. doi:10.1016/j.chb.2020.106548

14. Liu Z. Sociological perspectives on artificial intelligence: A typological reading. *Sociol Compass* (2021) 15(3):e12851. doi:10.1111/soc4.12851

15. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C. Machine behaviour. *Nature* (2019) 568(7753):477–86. doi:10.1038/s41586-019-1138-y

16. Jones GR, George JM. The experience and evolution of trust: Implications for cooperation and teamwork. *Acad Manage Rev* (1998) 23(3):531–46. doi:10.5465/amr.1998.926625

17. Beniger JR. *The control revolution: Technological and economic origins of the information society*. Cambridge, Mass: Harvard University Press (1986). p. 508.

18. Schilke O, Reimann M, Cook KS. Trust in social relations. *Annu Rev Sociol* (2021) 47(1):239–59. doi:10.1146/annurev-soc-082120-082850

19. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev* (1995) 20(3):709–34. doi:10.5465/amr.1995.9508080335

20. Lee JD, See KA. Trust in automation: Designing for appropriate reliance. *Hum Factors* (2004) 46(1):50–80. doi:10.1518/hfes.46.1.50.30392

21. Robinette P, Howard AM, Wagner AR. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Trans Hum Mach Syst* (2017) 47(4):425–36. doi:10.1109/thms.2017.2648849

22. Simmel G. *The philosophy of money*. London: Routledge (2004). p. 616.

23. Goffman E. *The presentation of self in everyday life*. Garden City, New York: Doubleday & Company (1959). p. 259.

24. Garfinkel H. *Studies in ethnomethodology*. Cambridge, UK: Polity (1991). p. 304.

25. Ward PR, Mamerow L, Meyer SB. Interpersonal trust across six Asia-Pacific countries: Testing and extending the 'high trust society' and 'low trust Society' theory. *Plos One* (2014) 9(4):e95555. doi:10.1371/journal.pone.0095555

26. Dickie J. *Cosa nostra: A history of the Sicilian mafia*. London: Hodder & Stoughton (2004). p. 483.

27. Gambetta D. *The Sicilian mafia: The business of private protection*. Cambridge, Mass: Harvard University Press (1993). p. 335.

28. Axelrod RM. *The evolution of cooperation*. New York: Basic Books (1984). p. 241.

29. Giddens A. *Modernity and self-identity: Self and society in the late modern age*. Stanford: Stanford University Press (1991). p. 256.

30. Zuboff S. *The age of the smart machine: the future of work and power*. New York: Basic Books (1988). p. 468.

31. Weber M. *Economy and society*. Cambridge, Mass: Harvard University Press (2019). p. 504.

32. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technol Forecast Soc Change* (2016) 105:105–20. doi:10.1016/j.techfore.2015.12.014

33. Latour B. *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press (2005). p. 301.

34. Grint K, Woolgar S. *The machine at work: Technology, work and organization*. Cambridge, UK: Blackwell Publishers (1997). p. 199.

35. Shestakofsky B. Working algorithms: Software automation and the future of work. *Work Occup* (2017) 44(4):376–423. doi:10.1177/0730888417726119

36. Jarrahi MH. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Bus Horiz* (2018) 61(4):577–86. doi:10.1016/j.bushor.2018.03.007

37. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: An attentional integration. *Hum Factors* (2010) 52(3):381–410. doi:10.1177/0018720810376055

38. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (2015). p. 1–8.

39. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* (2016) 3(1):205395171562251. doi:10.1177/2053951715622512

40. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag* (2019) 40(2):44–58. doi:10.1609/aimag.v40i2.2850

41. Ullman D, Malle BF. Human-Robot trust: Just a button press away. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction [Internet]. New York, NY, USA: Association for Computing Machinery.

42. Zlotowski J, Sumioka H, Nishio S, Glas D, Bartneck C, Ishiguro H. Persistence of the uncanny valley: The influence of repeated interactions and a robot's attitude on its perception. *Front Psychol* (2015). doi:10.3389/fpsyg.2015.00883

43. Simon H. Theories of bounded rationality. In: *Models of bounded rationality: Behavioral economics and business organization*. Cambridge, Mass: MIT Press (1982). p. 408–23.

44. Simon H. *Administrative behavior: A study of decision-making processes in administrative organizations*. New York: Free Press (1997). p. 368.

45. Rodrigues R. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *J Responsible Tech* (2020) 4:100005. doi:10.1016/j.jrt.2020.100005

46. Abbott A. Linked ecologies: States and universities as environments for professions. *Sociol Theor* (2005) 23(3):245–74. doi:10.1111/j.0735-2751.2005.00253.x

47. Pfeffer J. The role of the general manager in the new economy: Can we save people from technology dysfunctions? (2008). [Internet] 2018 [cited May 22, 2022] Stanford Graduate School of Business Working Paper No. 3714. Available from: https://www.gsb.stanford.edu/faculty-research/working-papers/role-general-manager-new-economy-can-we-save-people-technology.

48. Rafalow MH. Disciplining play: Digital youth culture as capital at school. *Am J Sociol* (2018) 123(5):1416–52. doi:10.1086/695766

49. Gusfield JR. *The culture of public problems: Drinking-driving and the symbolic order*. Chicago: University of Chicago Press (1984). p. 278.

50. Schlosser E. *Command and control: Nuclear weapons, the Damascus accident, and the illusion of safety*. New York: The Penguin Press (2013). p. 632.

51. Whitton J, Parry IM, Akiyoshi M, Lawless W. Conceptualizing a social sustainability framework for energy infrastructure decisions. *Energy Res Soc Sci* (2015) 8:127–38. doi:10.1016/j.erss.2015.05.010

52. Ishibashi K. Genpatsu shinsai: Hametsuwo sakeru tameni. *Kagaku* (1997) 67(10):720–4.

53. Stanley J. *Four problems with the ShotSpotter gunshot detection system*. News & Commentary [Internet]. New York: American Civil Liberties Union (2021).

54. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *Lancet Digit Health* (2021) 3(3):e195–203. doi:10.1016/s2589-7500(20)30292-2

55. Alexander JC. The societalization of social problems: Church pedophilia, phone hacking, and the financial crisis. *Am Sociol Rev* (2018) 83(6):1049–78. doi:10.1177/0003122418803376

56. Köchling A, Wehner MC. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Bus Res* (2020) 13(3):795–848. doi:10.1007/s40685-020-00134-w

57. O'Neil C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Reprint edition. New York: Crown (2016). p. 288.

58. Nowotny H. *AI we trust: Power, illusion and control of predictive algorithms*. Cambridge, UK: Polity (2021). p. 190.

59. Waller M, Waller P. *Why predictive algorithms are so risky for public sector bodies*. [Internet]. Rochester, NY: Social Science Research Network (2020).