The stuff we swim in: Regulation alone will not lead to justifiable trust in AI

Simon T. Powers, Olena Linnyk, Michael Guckert, Jennifer Hannig, Jeremy Pitt, Neil Urquhart, Aniko Ekart, Nils Gumpfer, The Anh Han, Peter R. Lewis, Stephen Marsh, Tim Weber

***Abstract*—Information technology is used ubiquitously and has become an integral part of everyday life. With the ever increasing pervasiveness and persuasiveness of Artificial Intelligence (AI), the function of socio-technical systems changes and must be considered as playing a more active role. Technology, e.g., in the form of large language models accessed through a chat interface, is now perceived as a social actor rather than as a passive instrument. Therefore, the question of how and when trust in technology and its organisational controllers is well placed is gaining relevance. In this article, we argue that simplistic views of trust that do not reflect the active nature of AI systems have to be replaced with more elaborate models. Regulation alone does not cover the complex relation between human user, AI system, creator, and auditor. We argue that a radical paradigm shift is urgently needed. The current debate that focuses the question of trust on explainable and ethical AI is dangerously misguided. Technology provides the opportunity for some organisations to leverage established prosocial trust relationships and repurpose them for their own narrow interests. The new model suggests an interpretation of socio-technical systems inspired by many-body physics, structuring interactions in a socio-technical system into fields and agents. This naturally explains the perceived agency of AI systems, and leads to actionable recommendations on how the discourse about trust can be reframed.**

***Index Terms*—Artificial Intelligence, Trust, Explainable AI, Regulation, Sociophysics.**

I. The New Age of Large Language Models

Recent activity in the field of AI has given rise to Large Language Models (LLMs) such as GPT-4 and Bard. These are undoubtedly impressive achievements, but they raise serious questions around appropriation, accuracy, explainability, accessibility, responsibility, and more. There have been pusillanimous and self-exculpating calls for a halt in development by senior researchers in the field, and largely self-serving comments by industry leaders around the potential of AI systems, good or bad. Many of these commentaries leverage misguided conceptions, in the popular imagination, of the competence of machine intelligence, based on some sorts of Frankenstein or Terminator-like fictions: however, this leaves it entirely unclear what exactly the relationship between human(ity) and AI, as represented by LLMs or what comes after, is or could be.

Most commentators would likely agree that LLMs represent a threshold change in the public's use of AI, but why is this? We argue that unlike previous AI technologies, such as deep learning classifiers performing image recognition, LLMs, especially when presented through a "chat" interface, are now perceived as social actors rather than passive instruments. Indeed, as more diverse capabilities are integrated alongside LLMs, the language of today's AI companies is shifting from talking of 'bots' to 'agents'[1], acknowledging this. The development and pervasive use of voice assistants, such as Alexa and Siri, over the last ten years has conditioned people to begin to anthropomorphise conversational AI systems [1]. But these early systems were very limited in what they could do and the responses that they could give. Moreover, they made it clear when a question was outside of their domain of competence, for example, by clearly stating that they were resorting to a web search to answer the question. LLMs, on the other hand, are able to (purport to) respond to questions on almost any topic imaginable. And the user can even customise the persona that the output of the LLM takes, e.g. GPT-4 can be prompted to answer in the manner of a lawyer, salesperson, or coach. The result is that LLMs have crossed the "uncanny valley" and afford interactions of the same style that someone might interact with a teacher, professional, or even friend.

The danger, then, is that people will use these systems for purposes far beyond that in which the systems are competent. For example, LLMs are not meant to provide factual information or professional advice, and typically contain a (small) disclaimer at the bottom of their chat interfaces indicating this. Perhaps unsurprisingly, some people nevertheless use LLMs for these purposes (e.g. in one study 78% of respondants were willing to use ChatGPT for medical self-diagnosis [2]). However, the Regulatory Theory of Social Influence (RTSI: [3]) posits that, in distributed information processing over social networks, not only are sources seeking potential targets to influence, but potential targets are also (for reasons of cognitive efficiency and coherence) actively seeking sources by whom to be influenced, putting their trust in those sources. The role of expertise in such situations is significant

Simon T. Powers and Neil Urquhart are with the School of Computing, Engineering and the Built Environment, Edinburgh Napier University. Olena Linnyk is with 1. milch & zucker Talent Acquisition & Talent Management Company AG, Giessen, Germany 2. Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany 3. Justus Liebig University of Giessen, Giessen, Germany. Michael Guckert, Jennifer Hannig and Nils Gumpfer are with the Department of MND - Mathematik, Naturwissenschaften und Datenverarbeitung, Technische Hochschule Mittelhessen - University of Applied Sciences. Jeremy Pitt is with Department of Electrical and Electronic Engineering, Imperial College London. Aniko Ekart is with School of Computer Science and Digital Technologies, Aston University. The Anh Han is with School of Computing, Engineering and Digital Technologies, Teesside University. Peter R. Lewis is the Canada Research Chair in Trustworthy Artificial Intelligence at Ontario Tech University. Stephen Marsh is with Faculty of Business and Information Technology Ontario Tech University. Tim Weber is with milch & zucker Talent Acquisition & Talent Management Company AG, Giessen, Germany

[1]E.g., https://www.gatesnotes.com/AI-agents.

[4], – or in this case, *apparent expertise*. By leveraging an existing form of social relationship, some systems, such as LLM chatbots integrated into web search engines, seem to encourage (unintentionally or otherwise) this misidentification and misappropriation in order to increase their user base.

This raises serious questions about under what circumstances people can trust these systems and how they can be developed in a trustworthy manner. Crucially, will they act in people's best interests [5]? Commentators often fall into two camps with their views on this. On the one hand, many commentators think that the solution is given by the magic bullets of regulation and explainability – regulators should ensure that AI systems are only released if they can explain why they have produced a certain output. On the other hand, some commentators are fatalistic and think that AI cannot be controlled and will lead to the demise of human society. In this paper, we argue that neither view is correct, because both rely on an inadequate model of human-AI interactions that lacks the social context. Specifically, the existing dialogue ignores the importance of the environmental context in which AI systems are developed and used, and hence in which users decide whether to trust them or not. This environmental context can be broken down into the state of

- culture,
- knowledge,
- nature,
- structural power,
- technology.

The effects of any regulations, or of interventions to make an AI system more "explainable", depend on these five environmental contexts as experienced by the user. They are the "stuff we swim in" all the time [6], and most of the time we are not consciously aware of them, even though they affect our decision making. We first draw on research at the interface of physics and social science (sociophysics [7]) to argue that the five contexts can be meaningfully described as 'social force fields', analogous to force fields in the physical sense. This physics formalism highlights a duality between fields and agents – when a background field, such as technology, is changing rapidly, it can be viewed as an agent in its own right. This field-agent model suggests that AI has led to technology becoming the most important field, and that this now drives the dynamics of our socio-technical society at the expense of the influence of the other fields. This implies that efforts to control the development and direction of AI technologies through regulation alone are unlikely to succeed. But on the flip side, the model suggests practical interventions that can be made to restore the strength of the other fields, in turn allowing for regulation to have its intended effects.

II. Triangle – Points in Space

The standard narrative of how the increasing use of AI-based data driven systems can be controlled is that regulation together with adequate transparency will prevent misuse and guarantee a contribution to general welfare. Before AI applications can be deployed regulatory bodies certify that relevant rules for creators and providers of AI systems as well as for applications are respected. Regulatory bodies, the AI system creator, and the AI system naturally form a triangle. Rules and certification are fundamental for human users to trust AI systems (see Fig. 1). This principal view is mirrored in current models of regulations for information systems, e.g. the General Data Protection Regulations (GDPR) [8] and the AI Act of the European Union [9]. For example, GDPR formulates a "right to an explanation" for human users. Systems that support automated decision-making have to provide an explanation for how a decision was derived. AI system creators are then forced to "care" about explainability and transparency as prerequisites of trustworthy AI. Furthermore, under the AI Act applications are categorised into four risk levels from unacceptable risk to minimal risk. Systems in the unacceptable-risk category, including social credit scoring systems, will not be permitted in the European Union. For high risk systems, e.g. most medical applications, a certification through a CE-marking process is mandatory. To sum up, regulation and transparency are supposed to be the key means of establishing trust in socio-technical systems containing AI systems. Note that to date the regulation systems are still work in progress and are not yet fully implemented.

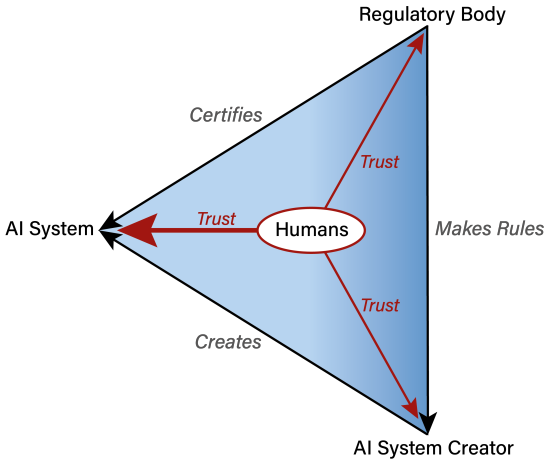

Fig. 1. Triangle – the traditional model of AI development governance, which includes regulatory bodies, AI system creators, and AI systems. Regulatory bodies make rules for AI system creators, and certify the resulting AI systems they produce. Users then trust an AI system because they trust the system creator and its regulators. Lewis and Marsh [10] describe this as 'proxy trust'.

III. Square – Introducing Fields

As set forth in the last section, the traditional narrative assumes that regulation accompanied by transparency will establish trust in the triangular relationship between human use, AI Systems, AI System creators and regulatory bodies and eventually lead to a proper, human-centred use of AI-based technology. However, while regulation obviously works beneficially in sufficiently static contexts – see the correlation in many countries between legally required, periodic inspection of vehicles and the number of road accidents [11] – there is also historic evidence indicating that regulation alone cannot be the magic bullet in dynamic environments. Agents pursuing objectives with high individual benefit will employ

changing circumstances for their advantage. The notorious CumEx scandal is an obvious example in which regulation could only detect loopholes when they were already being exploited (see [12]). Regulatory bodies have to respect manifold and contradictory interests and are therefore often too slow. In the hare and tortoise race between technological progress and regulation, directives lag behind and come too late. Moreover, they often finally overshoot and become obstacles for potential beneficial developments. This implicitly favours big tech companies that possess the requisite resources to stay ahead in innovation processes.

The complex relations in socio-technical systems are not as simple and static as the simplistic triangle would suggest, because integral factors such as structural power and knowledge are disregarded. These aspects were often not conceived as relevant in the past. However, with the World Wide Web – 1.0 as well as 2.0 – we have already experienced how commercialisation and aggregation of power have transformed a technological medium from its early innocent beginnings to what it is now. The hope that the Web 2.0 phenomenon was about democratising digital technology [13] has turned into a nightmare of fragmented epistemic universes, hosted on the platforms provided by big tech companies. We see how influential, opinion-forming social media platforms can be seized and then controlled by individuals with highly questionable intentions [14], [15]. By no means should the mistakes made in the past be repeated, and therefore different possible future trajectories must be adequately represented in dialogues about trust and regulation.

To do this, we can draw on theory from physics. Social sciences have successfully applied field theoretic ideas from physics to model interactions and interdependencies in the analyses of behaviour in social structures, such as social networks [16]. The application of field theory to model the behaviour of individuals in groups has already been suggested in the 30s of the last century by Kurt Lewin, in order to create a bigger picture of the many forces affecting an individual decision maker [17], [18]. Extending these ideas, we derive a new perspective on trust in socio-technical systems.

We now rearrange and extend the triangle from Fig. 1. In a first step away from the old, obsolete model we aggregate apparently homogeneous elements of culture, knowledge, nature, technology and structural power into mean fields that form the environment. These fields are the influential background against which relevant actors now form a square, with the human user with its various roles in the centre (see Fig. 2). The fish in the centre symbolises humans in their habitat actively as well as passively acting and interacting in their environment. Besides this role human users may also act as decision makers (the question mark symbol) or as representatives of power (the top hat symbol).

We need to recognise that the triangle is the incarnation of a fairy-tale model which assumes that trust of humans in AI technology and applications can simply be built on regulation and certification. This totally ignores the fact that system creators as well as regulators can have objectives of their own that do not necessarily have to be beneficial for the human individual. In Fig. 2, we therefore consider four additional goal seeking entities possessing agency. Widening the scope of the terms *AI System Creator* and *Regulatory Body* used before, we now explicitly conceive *Government and Regulators* and *(Powerful) Organisations* – subsuming tech companies as creators of AI systems – as actors. This mirrors the transformative role and the decisive influence they actually have. Mirrlees [19] provides an introduction to this form of power in today's world with a focus on 'Big Tech', introducing a conceptual framework for its analysis in terms of structural and relational power. While organisations create and deploy AI Systems regulatory bodies still have the responsibility to define rules and certify systems. *Commentariat* (media), i.e. basically the fourth estate, and *academia* have both played an important role in the past as independent and hopefully incorruptible authorities in social development. They must therefore necessarily be incorporated as active elements in a comprehensive model. Commentariat and academia should not be guided by vested interests and should drive knowledge creation and its communication. Their commenting on the use of AI, its opportunities and potential risks, constitutes an important contribution to the incorporation of technology into society. Unarguably, industry is playing an increasingly important role in basic research in AI, usurping positions that were traditionally held by universities and research institutions [20]. This observation is a call to academia to regain and strengthen its position so that AI research will not be completely dominated by few Big Tech companies.

With these modifications we gain a more realistic model for trust relationships in socio-technical systems and AI, that incorporates all relevant entities either as active actors – human beings or represented by human beings – or as fields that form the influential background. However, we realise that this view may still not suffice in the world of constant change that we are facing today.

IV. Pentagon – The Duality: fields and agents

In order to fully understand the decisions of humans in relation to technology today, we need to take into account the five fields described above, but also recognise that the technology field is itself now rapidly changing as a result of AI. This means that we need a way of modelling: 1. the interactions between individuals in society, 2. the effects of the fields (technology, culture etc. on these interactions), and 3. the dynamics of the fields themselves.

To do this, we can draw on modelling techniques from many-body physics, which are adept at modelling the local interactions between particles, the effects of fields (e.g. gravitational, or electromagnetic) on those interactions, and the dynamics of how the fields themselves actually arise from the local interactions. The key to this is to recognise a separation of timescales. Local interactions, between individuals in a society, or particles in a tube, happen at a rapid rate, causing the individuals to regularly update their own state and hence change their own behaviours. These local rapid interactions are traditionally modelled in physics using the Boltzmann equation, which tracks changes in the position and momentum of particles after they have interacted. Computational social

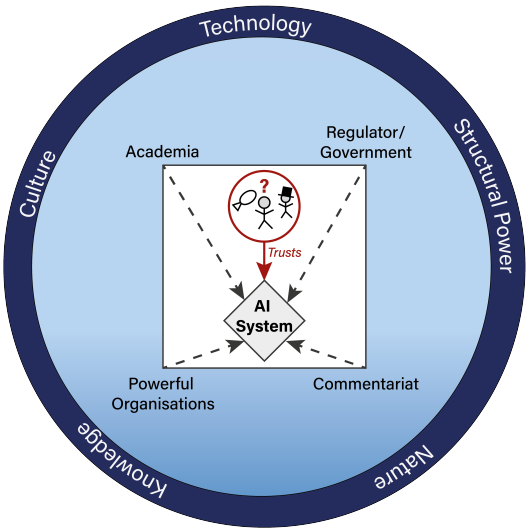

Fig. 2. Square – representing influential forces as fields. This improved model recognises that the extent to which an individual trusts an AI system depends on the background environmental factors of structural power, nature, knowledge, culture and the current state of technology. Moreover, the model also recognises that regulators and governments are themselves agents with their own goals, as are AI system creators, who are now labelled as powerful organisations to highlight this fact. Finally, the roles of academia and the commentariat as agents shaping public perception of AI are also recognised.

science models typically focus on these local interactions between individuals. They are, for example, at the core of Agent-Based Models (ABMs) [21], and of the replicator equation in Evolutionary Game Theory (EGT) [22].

In physics, the results of these interactions – the new position and momentum – are given not only by properties of the colliding particles, but also by the effects of the fields acting upon them. Likewise, how individuals in a society behave when they interact depends not just on their own knowledge and preferences, but also on the other fields (see Fig. 2). An example would be a social norm that is part of the culture field, e.g. "respect the word of elders". Such a norm would influence the direction of an individual's behaviour and the results of its interactions alongside the individual's own personal knowledge and preferences.

In physics, the effects of these fields on the interactions between individual particles are often modelled using mean field approximations. Mean field approximations treat the fields as constant with respect to the timescales over which particle interactions occur, allowing their effects to be analytically modelled using what is known as the Vlasov equation [23]. In computational social science, manifestations of the fields, such as social norms, can be treated as background parameters of the model that apply to all agents. For example, in models of indirect reciprocity, the same norm for assessing reputation is typically assumed to be shared by all individuals in the society (see e.g. [24]). The assumption is that the fields, such as social norms, change on a much slower timescale than individuals' decision making. Agent-based models that do this are effectively approximating the Vlasov equation.

However, when the fields are themselves changing rapidly, as indeed is the case with the popularisation of LLMs, this approximation is no longer valid. Technology is now moving too rapidly to be treated as a constant force field. We therefore now need a model that can incorporate both agent interactions and rapidly changing fields, and that is *self-consistent*. If we just simply had one set of equations to update the fields, and a separate set of equations to update the individuals' behaviour, then the model could be inconsistent, for individual behaviour may be updated based on an outdated version of the field, e.g. making an individual behavioural decision based on how the social norm was previously rather than how it is in the present time-step. Deriving a consistent model formalism for taking into account both effects – individual interactions and the fields – has to start from describing the dynamics of the system of many members (many-body-system) in terms of the hierarchy of correlations of $1,\ 2,\ 3, \ldots, N$-members. In physics, this leads us to arrive at generalised Boltzmann-Vlasov equations [25], [26].

Based on this, we propose that the Boltzmann-Vlasov equations can be used to formalise the *dynamics* of the model in Fig. 2. In order to be able to solve the equations, an effective theory approach [27], [28] from physics can be applied, in which the rapidly changing technology field is considered as the sum of a constant mean field and its fluctuation ($\phi = \bar{\phi}+\eta$). In this way, the effect of the changing field can be partly described by the interaction with effective degrees of freedom that are dual to the fields, i.e. by introducing additional agents into the simulation, which are generated by the fluctuations (η) in the mean fields ($\bar{\phi}$). These are not independent agents but rather "virtual" or "effective", generated by the field changes. This means that the technology field effectively becomes an agent itself in the model.

Fligstein and McAdam recently emphasised the importance of changes in social fields and suggested representing them by special field agents [29], [30]. They introduce "internal governance units" with the primary role of maintaining order, typically supporting the existing field state. They help stabilise fields, address crises, and connect with other fields. Examples of such units are certification boards in the professional fields and the World Bank. These can naturally be modelled the self-consistent description of the dynamics of fields and agents that we propose here.

Figure 3 shows this model conceptually. Compared to Fig. 2, this model captures the fact that AI technology is now rapidly changing. The AI system has now been brought inside the circle and become an agent in its own right, as formalised by the Boltzmann-Vlasov approach. This provides a formal model for the fact that many people do now view the likes of LLMs as agents in their own right. The model provides predictive leverage of the socio-technical trajectory of rapidly advancing AI, and suggests interventions that can be made to change this so that users are again given the agency to make an informed decision about to what extent they trust a particular AI system. We highlight some of these in the Actionable Recommendations section below.

The second change in Fig. 3 compared to the model in Fig. 2 is that the technology field has become much larger, and has compressed the other fields. This represents the fact that with rapid innovation in AI, technology is having more

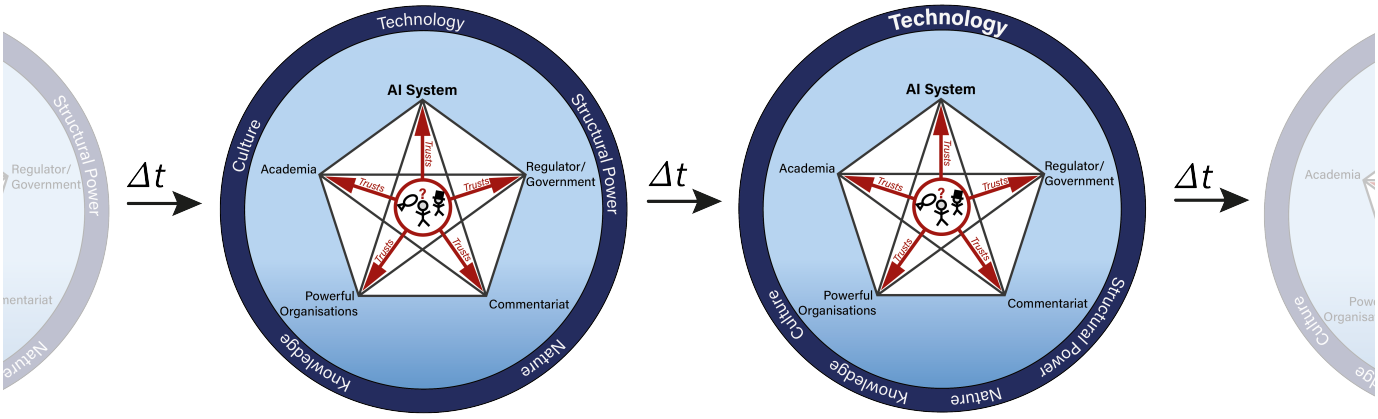

Fig. 3. Pentagon – The duality of AI systems as both field and agent. When a field (here technology) is rapidly changing, it needs to be modelled as an effective agent as well as a field (see text). The pentagon model therefore brings the AI system inside the circle as an agent, in addition to being a field. As time (t) progresses, the effect of this field-agent on the trajectory of the overall socio-technical system increases, compressing the other fields and leading to technology being the main driver at the expense of the other fields. For example, LLMs have come to dominate not only research but also education policy in universities and even the nature of future jobs.

of an effect than the other fields on individual behaviour, and is becoming the main driver of the socio-technical dynamics. Think, for example, of how in the last 12 months debates on education policy have been driven by the role of LLMs, and how university courses and assessment may need complete overhauls as a result. Likewise, structural power is now looking to be less manifest by governments, academia and the commentariat, and more by big tech, since only a select few big tech organisations can afford to train large-scale machine learning models such as LLMs (GPT-4, for instance, cost 100 million US dollars to train [31]). Yet while the model highlights this, it does not imply that this trajectory is inevitable. Rather, it suggests interventions that can be made to change this, which we discuss next.

V. Actionable Recommendations

In trying to understand and evaluate the societal impact of AI technologies, the transition from point to field to duality (and corresponding transition from mean-field background context to integral foreground actor, as per the Boltzmann-Vlasov equations) has revealed that (a) the situation is much more complicated than might have been thought; (b) it is much more deeply entangled than might have been thought; and (c) there is far less awareness of, and so appreciation of, the direction that societies are taking and how this will be felt by the individual – i.e., the very "stuff we swim in".

Therefore, there is an urgent need to educate citizens in order to enable them to engage knowledgeably and meaningfully in discourse about AI, and shape its trajectory. And here we do not only mean education in the traditional or formal sense, but in the broadest sense of AI or technology literacy; a comparison might be that while not everyone may formally learn to drive, a basic level of road safety literacy became essential for citizens following the widespread dissemination of the motor car. When considering Figs 1, 2 and 3, the outstanding feature is that the individual situated in the midst of the fields and is influenced, to some extent, by all of them. Education – in Information and Communication Technologies, ethics and citizenship – would provide people with at least a deeper awareness of the forces at work on them (even if they do not necessarily understand either the algorithms or the physics). Educating individuals also allows them to influence others through example, being themselves part of "the stuff we swim in" (cf. knowledge alignment in [32], social influence in [3]). But perhaps most importantly, education provides individuals with information that allows them to appreciate how the fields are changing right now. Once individuals have this information, they need not be mere passive recipients of the fields – particles being pushed and pulled by exogenous forces – but can actually *change* the fields by forming coalitions [33], [34]. It is import to stress that in this context we are not concerned with education of the technicalities of AI technologies (e.g. algorithms, statistics, data structures etc.) but on the more abstract issues around usage and ethics, and how these affect and are affected by the five fields, in particular the 'technology' field.

Within education we need to determine policies for teaching the use of AI and the ethics that surround that at an early age. Such education has to show the positive influences of AI as well as the negative ones, in order to avoid creating the impression that AI-driven movement through the socio-technical space in Fig. 3 inevitably leads to a worse position. It is also necessary that the adult population is encouraged into the AI discourse, in order to allow coalitions to change the fields and hence the trajectory through socio-technical space. There is a significant challenge in public education to ensure that the educational measures are accessible and influence all. There is a danger that educational measures may have a greater impact on those with more education and wealth and fail to empower others, leading to a divide within society; those who understand AI and can make a conscious decision as to when to allow themselves to be influenced by it and to form coalitions to shape the socio-technical trajectory, and those who are unknowingly influenced by AI and are so destined to

be mere recipients of the forces produced by the fields. One way in which this undesirable effect may happen is through the dissemination of incorrect or biased facts by LLMs, which people without access to this type of education may be inclined to view as agents rather than as fallible statistical models.

Measures are required that enable individuals to identify and protect their original creative works, versus those generated by AI. Educational measures must empower individuals to query the origins of creative works and understand the differences between one created by AI and one created by a human, and also to appreciate those works that may be termed hybrid having elements of AI and human creativity within them.

However, the role of AI in the co-creation of "works of mind" is, in fact, one particularly visible aspect of civic participation. There is a unique opportunity for using AI to support the development of next-generation public interest technologies [35], and re-empower civic participation in public consultations, deliberative assemblies, citizen science, data co-operatives, and other forms of local initiative. The pentagon model, though, highlights the agency of big tech organisations, and the fact that their goals are not necessarily aligned with the goals of the users of their produces. Indeed, some of these goals, intentional or not, seem to involve the reduction of users to aggregated revenue streams [36], or inhabitants of "Smart Cities" in which human personality is subordinated to technological rationality.

We cannot therefore rely on AI system creators (esp. businesses) to act ethically. Some system creators may label themselves as ethical, but in many cases these are likely to be the purveyors of open-source style technologies. Those AI system creators developing technologies such as LLMs have such high development costs that their subsequent actions with the technologies may prioritise income generating activities over ethical actions. It could be argued that regulatory measures could enforce AI system creators to behave in specific ways, but given the time that governments take to draft, approve and enact legislation there will frequently be a "law lag" between technologies being developed and regulatory measures being approved. In many cases the effectiveness of such measures may further rely on the legislation being enforced through a courts system and case law being established. There is also the risk of ethics "bluewashing", where developers make superficial or misleading claims about the ethical values of their systems [37]. Education of both the general public as well as other stakeholders such as executive board members can help guard against this [37], as can strengthening the role of the commentariat and academia.

Beyond civic education and civic participation, though, there are also more abstract dignitarian considerations [38]. Civic dignity, as a socially-constructed conceptual resource that facilitates collective action, is generated when citizens are treated as fully-fledged participants in deliberative democratic processes. It is, however, undermined when they are tricked into making decisions that, had they been fully appraised of the facts, they would not otherwise have made. Instead, we find AI being used for a range of human decision-making processes, from granting marriage licences to policing and sentencing, without indication, explanation, acknowledgement or, in particular, legitimate consent [39]. As Fig. 3 illustrates, as technology compresses knowledge, culture, and even structural power, and jurisdiction over "messy" human situations is supplanted by "unexplainable", let alone "justifiable" AI technology (there being a subtle but significant difference between giving reasons and proving reasonable), civic dignity can be rapidly diminished. This compels the question: at what point should citizens demand the right to a human decision [40]?

VI. Summary and Conclusions

In this article, we aimed to expand on the current discourse about the establishment of people's trust in AI and the development of AI systems that are trustworthy, to better reflect the need to be fully aware of what we otherwise take for granted – "the stuff we swim in". Building on the ideas brought forward at the first Trusting Intelligent Machines workshop [41], we have argued that the practical implementation of the conventional "triangular" model, and its extension to the "square model" are not just far from complete, they are not even fit-for-purpose for describing our rapidly dynamically changing world. We have argued that regulations alone are not sufficient: the much broader environment that we live in must be considered. While others have tried to understand this new "age of chaos" in general [42], here we have specifically proposed, for the first time, to model the trusting process using agents-and-fields theory (the "pentagon" model in Fig. 3).

In conclusion, though, perhaps it is the very definition of "AI" itself that needs to be re-examined. It is evident that the term "AI" is often used for marketing purposes, but has also been used to reinforce asymmetric power relationships, by leveraging misconceptions of computer capabilities and overblown fears of singularities in film, literature and the press [43]. Thus, from a historical technical perspective, the concept of "AI" has grown from relatively limited applications such as expert systems to significantly more complex systems such LLMs, open to a wide range of users and broadly applicable to numerous realms of human endeavour, and creating new disciplines such as 'prompt engineering'. In this process the definition of "AI" may become in some manner transient, focusing only on the latest technologies. It could therefore be argued that the recommendations proposed here should be applied more widely to software systems in general, and not just those which happen to have the label "AI" applied to them.

Moreover, we should not let the distractive opportunities of "AI" blind us to the (self-)destructive tendencies of a social, legal, political and economic system and its inability to respond appropriately to a number of inter-locking crises, for example in poverty, housing and climate change. It is one thing to host an "AI Safety" Summit as a vehicle for personal advancement or vanity, but ultimately it is not the "AI" that is safe, or unsafe: it is the design and intentions of the people, organisations and systems that produce, use and (supposedly) control it that need to be deemed safe, or otherwise (cf. [44]).

References

[1] G. Abercrombie, A. C. Curry, T. Dinkar, and Z. Talat, "Mirages: On Anthropomorphism in Dialogue Systems," May 2023, arXiv:2305.09800 [cs]. [Online]. Available: http://arxiv.org/abs/2305.09800

[2] Y. Shahsavar and A. Choudhury, "User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study," *JMIR Human Factors*, vol. 10, no. 1, p. e47564, May 2023, company: JMIR Human Factors Distributor: JMIR Human Factors Institution: JMIR Human Factors Label: JMIR Human Factors Publisher: JMIR Publications Inc., Toronto, Canada.

[3] A. Nowak, R. Vallacher, A. Rychwalska, M. Roszczynska-Kurasinska, K. Ziembowicz, M. Biesaga, and M. Kacprzyk, *Target in control: Social influence as distributed information processing*. Cham, CH: Springer, 2019.

[4] A. Mertzani, J. Pitt, A. Nowak, and T. Michalak, "Expertise, social influence, and knowledge aggregation in distributed information processing," *Artificial Life*, vol. 29, no. 1, pp. 37–65, 2023.

[5] T. A. Han, C. Perret, and S. T. Powers, "When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games," *Cognitive Systems Research*, vol. 68, pp. 111–124, Aug. 2021.

[6] D. F. Wallace, *This is water: Some thoughts, delivered on a significant occasion, about living a compassionate life*. Hachette UK, 2009.

[7] R. Kutner, M. Ausloos, D. Grech, T. Di Matteo, C. Schinckus, and H. Eugene Stanley, "Econophysics and sociophysics: Their milestones & challenges," *Physica A: Statistical Mechanics and its Applications*, vol. 516, pp. 240–253, Feb. 2019.

[8] European Parliament and Council of the European Union, *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence and amending certain union legislative acts*, 2021. [Online]. Available: https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en

[9] European Commission, *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence and amending certain union legislative acts*, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[10] P. R. Lewis and S. Marsh, "What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in Artificial Intelligence," *Cognitive Systems Research*, vol. 72, pp. 33–49, 2022.

[11] W. White, "Does periodic vehicle inspection prevent accidents?" *Accident Analysis & Prevention*, vol. 18, no. 1, pp. 51–62, 1986.

[12] C. Nikolaos. (2019) The cum ex scandal: financial crime and the loopholes in the current legal. [Online]. Available: https://policycommons.net/artifacts/2055068/european-parliament-p8_ta20180475-the-cum-ex-scandal/2808158/

[13] wiredstaff. (2006) Saving democracy with web 2.0. [Online]. Available: https://www.wired.com/2006/10/saving-democracy-with-web-2-0/

[14] H. STEVEN. (2022) Elon musk's twitter takeover is a colossally bad idea. [Online]. Available: https://www.ips-journal.eu/work-and-digitalisation/elon-musks-twitter-takeover-is-a-colossally-bad-idea-5912/

[15] P. R. Lewis, S. Lewis, S. Lewis, A. M. Gaudet, and A. Ottley, "Reimagining digital public spaces and artificial intelligence for deep cooperation," *IEEE Technology and Society Magazine*, vol. 42, no. 2, pp. 29–37, 2023.

[16] A. P. Alodjants, A. Y. Bazhenov, A. Y. Khrennikov, and et al., "Mean-field theory of social laser," *Scientific Reports*, vol. 12, p. 8566, 2022.

[17] K. Lewin, "The conceptual representation and measurement of psychological forces," *Contributions to Psychological Theory*, vol. 1, no. 4, p. 247, 1938.

[18] ——, *Field Theory of Social Science: Selected Theoretical Papers*, D. Cartwright, Ed. New York: Harper & Brothers, 1951.

[19] T. Mirrlees, "Getting at GAFAM's power: A structural and relational framework," *Heliotrope*, February 2021. [Online]. Available: https://geliotropejournal.net/helio/gafams-power-in-society

[20] R. Jurowetzki, D. Hain, J. Mateos-Garcia, and K. Stathoulopoulos, "The privatization of AI research(-ers): Causes and potential consequences – from university-industry interaction to public research brain-drain?" 2021.

[21] C. Adami, J. Schossau, and A. Hintze, "Evolutionary game theory using agent-based methods," *Physics of life reviews*, vol. 19, pp. 1–26, 2016.

[22] J. Hofbauer and K. Sigmund, *Evolutionary games and population dynamics*. Cambridge university press, 1998.

[23] A. Vlasov, *Many-particle Theory and Its Application to Plasma*, ser. Russian Monographs. Gordon and Breach, 1961.

[24] M. A. Nowak and K. Sigmund, "Evolution of indirect reciprocity," *Nature*, vol. 437, no. 7063, pp. 1291–1298, Oct. 2005.

[25] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics*. New York: Benjamin, 1962.

[26] M. Burger, "Network structured kinetic models of social interactions," *Vietnam J Math*, vol. 49, no. 3, pp. 937–956, 2021.

[27] S. Weinberg, *The Quantum Theory of Fields, Volumes 1-3:*. Cambridge University Press, 2005.

[28] J. Schwinger, *Particles, sources, and fields*. Reading, Mass.: Advanced Book Program, Perseus Books, 1998.

[29] N. Fligstein and D. McAdam, *A theory of fields*. Oxford: Oxford University Press, 2012.

[30] D. N. Kluttz and N. Fligstein, *Varieties of Sociological Field Theory 10.1 Introduction*. Springer, July 2016.

[31] W. Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, 2023, section: tags. [Online]. Available: https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[32] J. Ober, *Democracy and knowledge*. Princeton Univ. Press, 2008.

[33] J. Pitt, "Interactional justice and self-governance of open self-organising systems," in *11th IEEE International Conference SASO*, 2017, pp. 31–40.

[34] C. Perret, S. T. Powers, J. Pitt, and E. Hart, "Can justice be fair when it is blind? how social network structures can promote or prevent the evolution of despotism," in *2018 Conference on Artificial Life, ALIFE 2018, Tokyo, Japan, July 23-27, 2018*, T. Ikegami, N. Virgo, O. Witkowski, M. Oka, R. Suzuki, and H. Iizuka, Eds. MIT Press, 2018, pp. 288–295.

[35] R. Abbas, S. Hamdoun, J. Abu-Ghazaleh, N. Chhetri, N. Chhetri, and K. Michael, "Co-designing the future with public interest technology," *IEEE Technology and Society Magazine*, vol. 40, no. 3, pp. 10–15, 2021.

[36] S. Zuboff, "Big other: surveillance capitalism and the prospects of an information civilization," *Journal of Information Technology*, vol. 30, pp. 75–89, 2015.

[37] L. Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical," *Philosophy & Technology*, vol. 32, no. 2, pp. 185–193, Jun. 2019.

[38] J. Ober, *Demopolis: Democracy before liberalism in theory and practice*. Cambridge University Press, 2017.

[39] J. Pitt, "The principles of cyber–anarcho–socialism," *IEEE Technology and Society Magazine*, vol. 41, no. 1, pp. 5–10, 2022.

[40] J. Tasioulas, "Artificial intelligence, ethics, and a right to a human decision," Harold T. Shapiro Lecture on Ethics, Science, and Technology, Princeton University, 2023.

[41] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, N. Urquhart, and S. Wells, "Trusting intelligent machines: Deepening trust within socio-technical systems," *IEEE Technology and Society Magazine*, vol. 37, no. 4, pp. 76–83, 2018.

[42] J. Cascio, "Facing the age of chaos," *Medium*, April 2020. [Online]. Available: https://medium.com/@cascio/facing-the-age-of-chaos-b00687b1f51d

[43] P. R. Lewis, S. Marsh, and J. Pitt, "AI vs "AI"': Synthetic minds or speech acts," *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 6–13, 2021.

[44] C. Perakslis, R. Abbas, K. Michael, M. Michael, and J. Pitt, *Safeguarding the Guardians to Safeguard the Bio-economy and Mitigate Social Injustices*. Springer, 2023.