DOI:
10.1109/MTS.2023.3340232

Document Version
Peer reviewed version

Department: Head
Editor: Name, xxxx@email

Deceptive AI & Society

Ștefan Sarkadi
Dept. of Informatics, King's College London, UK

***Abstract*—'Deceptive AI' is an expression that captures the imagination of both users of AI technologies, and AI experts or researchers alike. But what does 'Deceptive AI' mean? Figuring out the meaning of this term is important to understand not just the recent hype around language models and their implementation in chatbot technologies, but also the historical gravitas behind building machines capable of deception, as well as how to take things forward from here. 'Deceptive AI' can refer to a multitude of different things, yet its entire range of meanings is intertwined with the human condition, i.e. how humans represent their own selves in the world, and the idea of hybrid societies, which are societies where humans and machines play the role of agents in a society. In this paper, I will take the reader through the various aspects that are relevant to deceptive AI research and try to explain the relations between the theories and technologies behind deceptive AI technologies, human-like intelligence, society, and the ethical principles and potential regulatory mechanisms that apply to deceptive AI technologies.**

DECEPTIVE AI is a heavily loaded term. Its semantic load has become exponentially heavier in a very short period of time. Perhaps most of this semantic load, at least in the recent public sphere, has been placed on it because of the deployment of large language models, such as ChatGPT. Deceptive AI is very multi-faceted. There are different AI approaches which give rise to different types of AI technologies, or in some case autonomous agents. Some of these technologies already exist in practice, others exist in theory, some are transitioning between theory to implementation, and, finally, some are still only fictions of our shared imagination [62].

All of these AI technologies can be deceptive in their own way [40]. Some are intrinsically deceptive, and others are extrinsically deceptive. Some are mere tools used for deception by humans, and others have their own deceptive goals. Some are designed to be deceptive, others are endowed with the capability to deliberate whether they should pursue deception or not. Some are perceived by others to be deceptive, when in fact they do not 'intend' to deceive, while others might never be perceived to be deceptive, yet the ulterior goals they pursue makes them perform on over-arching and sophisticated deception that might never be revealed by others.

The deceptiveness of AI technologies is strongly dependent on the human condition. By human condition I mean the way in which humans represent their own selves in the world, where the world consists of their own embodiment, the environment, the agents that share their environment, and the rules they follow to act in that shared environment, i.e. the society they are

 Published by the IEEE Computer Society

part of. As humans, we resort to anthropomorphic tendencies when we explain things. Explainable AI is an active research area that emphasises the explanatory power of anthropocentric concepts, such as 'intention', 'belief', 'goal', 'desire', 'knowledge', 'plan', 'other', 'self', etc. [42, 39]. Explanation here plays a key role regarding how we explain deception and what we consider deceptive AI to be [58]. Most research on deception comes from Communication Theory [37], where social and psychological factors involved in deception and deception detection are used to build theories that explain these phenomena in humans. In developmental Psychology, deception is also considered to play a huge role in how humans actually develop their conceptual understanding, self-representation and representation of others through exploitation and understanding of false beliefs [21, 69].

As Castelfranchi once predicted, machines have now come to deceive us and each other [8]. But we have also come to deceive machines ourselves, use machines to deceive others, and even use machines to deceive ourselves. This is not just in single agent to single agent interactions, but also in multi-agent to multi-agent interactions [17]. Beyond the business hype and the tendency to assign human-like properties to cognitively poor systems like ChatGPT [71], there are actually more dangerous activities enabled by deceptive AI technologies that pervade our societies, such as psychological warfare [67, 64].

The concept of deceptive AI is intertwined with multiple facets of our human condition. To be able to describe what the term implies is impossible without referring to society and how the idea of AI has co-evolved with society. Hence, to begin describing this co-evolution, we must first go back to the inception of deceptive AI in the history of computing.

BACK TO THE FUTURE

When did deceptive AI start? For most AI researchers it is rather obvious when, namely along with Turing's notorious *Imitation Game* (a.k.a. The Turing Test) [68]. Turing introduced the subtle concept of a machine capable of tricking humans into assigning them the property of intelligence.

The mere idea of such a test fuels a machine's deceptiveness, as it does not actually evaluate the existence of a *mind* behind the observed behaviour [27]. Unfortunately, the Turing Test has been misinterpreted as a 'pen-pal' one-shot interaction, rather than a life-long evaluation of the presence of an intelligent mind [28]. Subsequently, this interpretation led researchers to optimise for AI's trickery abilities, rather than modelling the internal cognitive processes which would enable the genuine and deliberate ability to trick others about one's type of embodiment (machine or human) [29].

However, throughout its history, AI research has not forgotten its original goals, those of making *'computers do the same things human minds can do'* [4, p1]. While Turing's game was interpreted by some as to create AI technologies that use trickeries to trick judges instead of creating the minds for the AI to be able to trick judges, others have thought about creating artificial minds that are able to have a meaningful dialogue with the interlocutor. One such line of work was initiated by Hamblin on constructing mathematical models of dialogue that are based on pragmatics [66, 25][1]. Another line of work arose from Cybernetics with Pask's Conversation Theory that propositions the idea that knowledge discovery, consolidation, and concept formation happens through conversation [48]. Both Hamblin's and Pask's ideas have something fundamental in common, namely that what matters respective to a dialogue or conversations happens inside the mind. For instance, both Hamblin and Pask refer indirectly to machines capable of using a Theory of Mind (ToM) to reason about the content (semantics) of their utterances. Hamblin calls it a dynamic knowledge base of interlocutor's beliefs, while Pask calls this a 'Cognitive Reflector'. No such modules are being referenced in the current data-oriented approaches used to create 'stochastic parrots' [3], which follow the trend of creating AI trickeries for tricking, rather than AI minds capable of tricking.

DECEPTIVE AI SYSTEMS

We could playfully name the two main approaches to Deceptive AI as 'truly Deceptive

[1] After Hamblin's death, Staines edited Hamblin's manuscripts as the book entitled *'Linguistics and The Parts of The Mind'* [66].

AI', and 'fake Deceptive AI', where the former approach is concerned with creating the cognitive processes and architectures that would enable deception in dialogue, whereas the latter approach is concerned with building 'mindless' machines that have been behaviourally 'trained' or conditioned through correlation to engage in dialogue 'as if' they have a human-like mind or a subset of human-like cognitive capabilities that enable them to do it.

One thing is to distinguish between approaches used to create deceptive AI, another is to differentiate between the types of AI that are potentially deceptive.

If we are to start by categorising the type of deceptive behaviour, then we need look no further than Masters et al.'s characterisation of deception in AI systems [40]. They distinguish between five different types of deceptive behavioural patterns that each have different consequences on humans, namely imitating, obfuscating, tricking, calculating, and reframing. Imitation and tricking have been achieved through generative AI and data-oriented approaches, e.g. DeepFakes, false statements etc. Obfuscation, a behaviour present in human military scenarios, has been achieved in AI through deceptive path-planning. Calculating deception has been studied in game-theoretical AI approaches and is performed by exploiting the partial information available to the target. Finally, reframing relies on establishing a complex pattern of behaviour that feeds a prior false belief or set of beliefs in the mind of the target. Yet, this type of characterisation can just as well be applied to humans.

But what about what's happening in the 'mind' of deceptive AI models? Another distinction we can make between different AI systems is between AI tools, fully autonomous AI agents, and everything in-between.

From Deceptive Tools to Deceptive Agents

Current deep learning techniques that underlie software such as DeepFake [57] or language models that underlie DeepFake Text [53] offer the possibility of creating deceptive digital content. However, these techniques do not offer an AI architecture that is in itself capable of deception, i.e. that is able to reason about the minds of others and to decide what information should be used to manipulate others' beliefs. Fortunately, there is no artificial human-like mind behind these models that decides what type of information needs to be distributed online such that it deceives web users, not even behind LLMs such as ChatGPT [3]. That does not mean that human-like deceptive machines cannot be engineered and deployed.

There are already implemented models of artificial agents that distinguish between the different ways of representing and communicating a lie, telling nonsense or even to employ one-way deception if they have a model of their target's mind [47]. It is easily conceivable to imagine the possibility for such agents to either be given a Theory of Mind (ToM) of their targets or the ability to form a ToM of their targets and reason about it abductively. The application of ToM in multi-agent systems (MAS) is an active area of agent-oriented AI research [56, 50]. Its application in deploying deceptive AI would enable a Hamblin and Pask style of pragmatics-oriented communication process that goes beyond the correlational conditioning of LLMs ('stochastic parroting'). Subsequently, we can easily imagine a strategic deployment of these AI agents on the internet, not as tools directed by humans, but as fully autonomous deceptive agents.

For an AI agent to be capable of deliberate deception [58] it must be capable of metacognition [13], i.e. it has to be able to know or to believe things, to know what it knows or believes, to know what it doesn't know, but also to know what its target does not know what the target does not know (nested reasoning), all in a recursive manner as modelled, albeit on a very high-level, in [60]. The agent has to be able to reason up-and-down its own representation of itself and the world, similar to going up-and-down the levels of Dennett's tower of cognition [15]. Such an AI agent must have several *self-** properties, such as self-awareness and self-adaptation (or self-expression), but also representation of self-awareness of other agents that share its environment and or social system, i.e. a Theory of Mind of self-aware agents.

The literature on self-aware systems emphasises the importance of different levels of self-awareness, self-adaptation of computing architectures and evolutionary interactions in AI [34]. Another important aspect in self-aware AI is

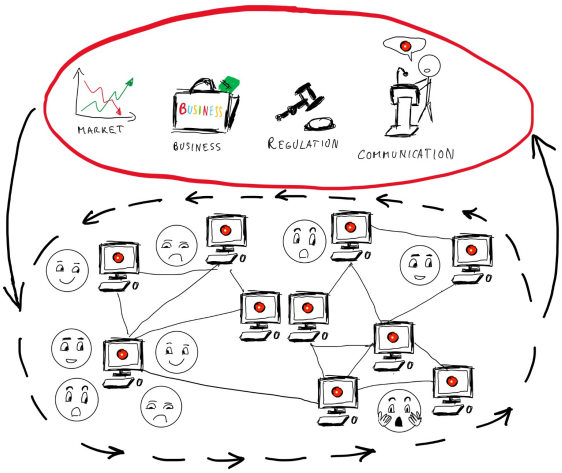

Figure 1. A deceptive AI ecosystem [71] represents not just the technical aspects of developing deceptive AI technologies, but how the societal and evolutionary pressures influence human interaction with these technologies as both individuals, groups, and organisations. This creates an ever evolving informational feedback loop between hybrid societies where humans and machines communicate as agents and the emerging socio-economical regulatory norms, human and societal values, business decisions, power structures, communication about AI technologies and market behaviour.

that of modelling cognitive processes and how this enables us to make self-aware AI agents predictable and explainable without having to explicitly model their domain of operation at runtime or a priori [38], e.g. by applying epistemic planning [6]. In the case of deception, this level of self-awareness would be a meta self-awareness [45, 2].

The modelling of cognitive AI agents with self-* properties, helps us not only understand, predict and explain the AI agents' behaviours, but also their internal decision-making properties. These internal cognitive processes can then be used to inform the analysis of emergent strategies and their evolution over time driven by interactions between populations of agents. Now we go to a higher level of abstraction, where another self-* property plays a crucial role in analysing, and perhaps even regulating deception, namely that of self-organisation [51].

DECEPTIVE ECOSYSTEMS

In the area of multi-agent systems, self-organising agent populations can also represent hybrid societies that consists of both humans and machines. Today's societies are becoming increasingly hybrid, and this increased hybridisation brings with it both benefits and risks. In the Philosophy of Information, the long-term negative impact of deceptive AI on hybrid societies is described by Greco and Floridi [22], who introduce the notion of *The Tragedy of the Digital Commons* (TDC), which is an extension of Hardin's concept of *Tragedy of The Commons*, where self-interested individual human agents are joined by artificial agents who also exploit a common good/resource to the point where the common good cannot be replenished due to over-exploitation [26].

AI agents can misuse and 'pollute' either (i) through exploitation such as extensive generation of information like spam or self-replication of a computer worms (which also consume bandwidth, and, in turn, this restricts access to information), or (ii) through destruction, such as the deletion of information from database systems.

This perspective encourages us to think of deceptive machine behaviour as part of an **ecosystem** (see Fig. 1), rather than as an isolated event between two or more agents [54, 71]. Deceptive AI can thus impact the evolution and stability of a multi-agent ecosystem. Indeed, such a free-riding effect is studied by Sarkadi et al. [61] that used extensive self-organising multi-agent simulations to demonstrate how deceptive agents (humans or machines) could break down cooperation in hybrid societies and cause a TDC. But how would this look like in the real world?

In the area of **cybersecurity**, deceptive AI has usually taken the form of online troll-bots [36] and of particular cases of social engineering that revolve around accessing sensitive computer data [43]. Social engineering attacks are already being scaled up through the use of ChatGPT, e.g. by prompting it to write persuasive emails [23]. A future potential threat to security would be the full automation of AI agents to be able to employ social engineering attacks in order to reach their own malicious goals. For example, we could imagine a fully autonomous AI agent that

not only knows how to write a computer virus to destabilise a country's critical national infrastructure, but also how to manipulate other machines or humans to use this virus or to carry the virus it to its destination, similarly to a Stuxnet [35] virus that is able to deliver itself onto secure isolated networks through social engineering.

Regarding the **erosion of trust and democracy**, we can also imagine a mastermind deceiver machine that manipulates others to propagate the machine mastermind's lies or fake narratives through social networks of agents. Going even further beyond the possibility of software that can generate fake media and of agents that can deceive and coerce, we can infer the possibility of autonomous fake news agencies, or AI troll farms run by AI agents themselves, in a distributed and decentralised manner. Human organisations or individuals could potentially hire these agencies to perform certain tasks. They could, for example, give an autonomous fake news agency the goal to increase someone's popularity, as is currently the case with troll farms used for propagating disinformation [1]. The agency would then gather data on its own to form the ToMs of its target audience, and then would plan what information to forge (or not) and what information to disseminate in order to achieve its goal. This scenario is a threat to accountability. Humans can be held responsible for unethical behaviour, but how are we going to hold artificial agents responsible for the creation and dissemination of not only fake news, but of massive deception operations, especially when they may be able to out-reason us or the machines that are aligned with us?

In the **marketing and entertainment** domains, the advancement on embodied and emotionally intelligent artificial agents [49, 33] could play a major role in the manipulation of affective social interactions. For example, empathic deceptive agents might simulate the emotional states through facial expressions or other physiological cues of a trustworthy AI in order to increase their target's trust, in a way that is similar to the way that psychopathic human agents mimic the emotional responses of non-psychopathic agents [52]. The ability of machines to feign emotions can have an impact on their targets' perception and biases, hence influencing their targets' opinions. In some contexts feigning of emotions during everyday social interactions can be considered benign, while in other contexts, it might have serious implications especially if they could impact critical decision-making in a systemic manner [18].

In the **legal** context, there are also problems that can emerge from the ability of machines to argue and build stories to use in court. This could be quite a serious threat, as LLMs have already been used to build legal cases, even if poorly so [5]. Will the future see deceptive AI agents hired to defend human criminals, or even machine criminals, from being held responsible? Let us assume that we would be able to develop a method for holding machines responsible for the unethical behaviour along with a legal system in place that would allow prosecutors (human or machine or both) to interrogate and analyse deceptive machines. What if the deceptive machines are able to hire their own lawyers or even to pay engineers to extend their architecture such that they are able to defend themselves in a legal manner? What would the combination of a deceptive agent architecture with such an ability imply? In hybrid societies this phenomenon could actually trigger an **arms race** in Theory of Mind as envisioned by Dennett [14], and then modelled and shown in simulations by Sarkadi [59].

PRO-SOCIAL DECEPTIVE AI

On a more positive note, echoing Ostrom's ideas for governing the commons, Sarkadi et al. [61] also showed that if the right regulatory mechanisms are in place for social interaction, cooperation could be re-established when deceptive agents are present and a TDC could be avoided. The solution found by Sarkadi et al.'s model was the promotion of a governing mechanism that is able not only to punish malicious deceivers, but to do so through a thorough process of interrogation, clarification of evidence, and debate in a decentralised manner. Their solution to TDC echoed Habermas's ideas in that of promoting a Digital Public Sphere where debate and interaction would happen in hybrid societies. There is a catch though, when ToM capabilities start emerging, investigators have to be able to out-reason the deceivers [63]. Still, the question remains: *how does such a regulatory mechanism look like in the real world?*

Ethical Guidelines and Deceptive AI

Either reaching a TDC due to deceptive AI agents or mindlessly implementing deception-based algorithms in society brings us to the necessity of understanding how to regulate deceptive AI. However, to be able to do so, it is necessary to know what we are regulating for or against. A starting point for designing regulation can be the ART principles of Trustworthy AI proposed by Dignum, namely accountability, responsibility and transparency for ensuring an ethical design of AI systems [16]. Expanding on these principles, the EU High-Level Expert Group in AI specified a set of guidelines [30].

Regarding deception, the EU's guidelines explicitly point out the perils of deceptive technologies w.r.t. fundamental rights as a basis for Trustworthy AI. Specifically, the right to *freedom of the individual* can be violated by AI through deception and manipulation (or even coercion due to deceptive design) of humans into making decisions that humans would otherwise not make. This type of AI behaviour undermines human agency, that is a human's ability to make informed autonomous decisions regarding AI systems. These guidelines explicitly emphasise the need to adhere to *the principle of respect for human autonomy* in the context of building AI systems.

Implicitly, the EU's guidelines point out three other ethical principles in the context of AI systems, which indirectly refer to deception. These are (i) the principle of prevention of harm; (ii) the principle of fairness; and (iii) the principle of explicability.

In their Ethically Aligned Designed (Version 2) report [10], IEEE proposes a set of principles that are in tune with the EU's. For instance, the idea that AI systems must be both transparent and 'honest' in order to promote non-deception. Additionally, the report emphasises the potential risks, such as the use of deception to cause even well designed autonomous weapons act against an incorrect target. Both risks and benefits are mentioned in the report in the context of affective computing, where AI systems might manipulate or deceive for both their own and others' benefit, e.g. nudge a human to do the right thing [7].

So, can deceptive AI technologies actually promote the principles ingrained in the EU's and IEEE's guidelines, or are deceptive technologies doomed to only pose risks to hybrid societies? We will look at two arguments that explain how this can be achieved.

The Ethically Aligned Deceptive AI Argument

The first argument is that the ability to deceive is crucial for machines to be able to interact socially in a smooth and meaningful manner. Isaac and Bridewell [31] describe this argument by introducing the perspective that in some circumstances, machine deception is socially beneficial.

However, Isaac and Bridewell also argue that the condition for a deceptive machine to be ethical is to be ethically aligned with its target (the one being deceived). This alignment means that the machine is able to reason about the ethical and moral values its target holds. Hence, in order to deceive in an ethical manner, machines must distinguish between morally permissible and impermissible ulterior goals. For instance, if the target believes that deception is never ethical, then the machine will reason that it is impermissible to attempt deception with that target and will drop pursuing its ulterior goal that constrains it to deceive in that particular circumstance. Counterfactually, if the target believes that deception is beneficial when used to avoid further pain (physical or psychological), then the machine might reason that it is ethically permissible to attempt deception with that target in specific circumstances and will follow its ulterior goal.

A more tangible example is that of a machine in the role of a medical doctor that interacts with a terminally ill patient. The machine has the option to communicate to the patient the diagnosis, yet it knows from the previous interactions with the patient that the patient would prefer not to know that they are terminally ill, as this would then have negative effects on their mental health and life experience. The machine then has the option to deceive the patient, as this deception is aligned with the patient's moral values.

Furthermore, the authors argue that in order for a machine to be able to align itself correctly about another agent's ethical values and to reason in terms of ulterior goals, it requires a model of the target's mind. Conclusively, deceptive machines can be beneficial if they are able to reason about the mind of their targets in order

to distinguish between the ethically permissible and ethically impermissible acts from the target's perspective. Moreover, for these machines to be trustworthy, they must also be able to provide explanations for why or why not they attempted deception in order to justify their decisions [39].

Ethically Aligned Persuasive AI

AI-powered machines are not so advanced yet as to be responsible for forming their own models of their interlocutors and making ethical decisions based on these models in the same way Isaac and Bridewell suggest they do. Yet, the same argument, but under a specific context, namely that of entertainment, is made by Coeckelbergh, who argues that while deception as a co-performance is beneficial for entertainment, there must be some contextual safeguarding behaviour [11]. For instance, if a machine is performing an illusion to entertain the human, then the designers of the machine should ensure that the context in which the illusion is performed is well defined and that the human is made aware of it. Once the trick is performed and the human was entertained, the human must be informed that it was only a temporary illusion for the sole purpose of entertainment. Now, differently from the previous argument where the designers of the agent must only ensure that the machine is capable of aligning itself with the interlocutor, Coeckelbergh's argument also puts the burden of alignment on the designers regarding the technological context.

A specific area of entertainment where this is applied is gaming, where virtual characters are designed such that humans can easily anthropomorphise them, and are incentivised to do so as part of an immersive form of narration. Another example from real world domain where deception and AI meet is marketing. Marketing techniques that have always relied on persuasion [41] are now enhanced by advancements in HCI on persuasive technologies [19]. One example is the design of digital platforms or applications to use deception and coercion for persuading their users [32].

One could argue that most of these 'persuasive' technologies are actually 'deceptive' technologies, and that in most cases, the deceptive design behind these technologies exploits the human truth-default state. Natale et al. call this *banal deception* [44].

Additionally, one can draw similarities between banal deception and the way Coeckelbergh describes deception as a phenomenon co-created and co-performed by humans or robots [11], e.g. using AI for stage magic. Notably, Coeckelbergh emphasises the need for this type of deception (as a co-performance) to be open, consensual and transparent, in order to achieve constructive persuasion [12].

In other words, the design of deceptive technologies for the purpose of entertainment could very well follow the ART principles and EU guidelines to ensure that their co-performance prevents harm, is fair, and explainable.

The Scientific Discovery Argument in AI

The second argument for pro-social deceptive AI considers the creation of Hamblin style AI agents. So, why should we research how to design Hamblin style AI agents if they risk becoming deceptive?

To reason about this, Sarkadi proposes the adoption of Resnick's method of *Reflective Equilibrium* [55] in the context of scientific discovery and machine deception in order to use (i) unbiased, reflective judgments or intuitions about what is or what would be considered right or wrong in particular contexts, e.g., the context of modelling deceptive machines; and to (ii) propose theories and principles which aim to provide a coherent justification of these judgments [58]. As a first step, the following general ethical principles introduced by Resnick can be applied:

1) The non-maleficence principle: One should not act in ways that cause needless injury or harm to others.
2) The beneficence principle: One should act in ways that promote the welfare of others.
3) The intellectual freedom principle: One should be allowed to pursue novel ideas and criticise old ones. One should be free to conduct research they find interesting.
4) The openness principle: One should allow people to see their work, and be open to criticism.
5) The honesty principle: One should aim towards finding the truth and should communicate in a truthful manner.

The second step is to propose an ethical argument that can be made **against** the scientific exploration of modelling deceptive machines. One such argument is that deception is unethical because this scientific exploration might lead to the development of fully autonomous deceptive agents that will deceive humans. Therefore we should not try to model deceptive machines. An elaborate account of this line of reasoning can be found in the controversial AI-Box Experiment in which a super-intelligent and malicious AI agent that is locked inside a software sandbox (a virtual prison) deceives a human, the guard of the box, in order to escape form the box and wreak havok in society [70].

With respect to the five principles enumerated above, a **counter-argument** can be made to support the ethical modelling of Hamblin style deceptive machines. By modelling such deceptive machines in multi-agent systems, we are able to understand them, e.g., their internal mechanisms as well as how they might interact with other agents in complex social systems. W.r.t., the **non-maleficence principle**, this understanding might prevent us from actually creating or enabling deceptive machines to act in ways that can cause harm to others. W.r.t., the **beneficence principle**, we could understand how deceptive machines might be created such that they provide benefits to society, i.e. deceive in an ethical manner to achieve an ulterior goal (see [31, 65, 9, 63]). W.r.t., the **intellectual freedom principle**, researchers can personally find that the topic of deception, and machine deception in particular, are simply fascinating because 1) deception has a certain historical gravitas in the area of AI given its exploration in Turing's Immitation Game as a necessary requirement for humans to assign the property of intelligence to machines [68], and thus being central to the antropomorphisation of artificial agents (another fascinating topic in itself); 2) deception is a popular recurrent topic in science-fiction that when consumed by the general public it forms the public opinion of AI (see [20]); and 3) deception is a very complex phenomenon in terms of its psychological, evolutionary and epistemic properties, and the idea of modelling these properties in interactions between artificial agents is very exciting from a scientific perspective. W.r.t., the **openness principle**, the actual modelling of deception in a public and scientific way promotes this principle by opening a much needed well-informed discussion about the topic that goes beyond anecdotal explorations and that can better inform public opinion. W.r.t., the **honesty principle**, modelling deception in order to better understand it and sharing this understanding in an honest manner is a truth-promoting act in itself, independent of the ulterior motives of performing the act.

This line of reasoning supports the idea that research on creating Hamblin style deceptive AI is ethical if it is applied to promote and is guided by ethical principles which enable scientific discovery.

CONCLUSIONS

Our societies are being challenged by a multitude of problems due to deceptive AI. To understand the roots of these issues, in this paper I aimed to explain the many facets of deceptive AI, i.e. its meanings. These facets are historical (the goals of deceptive AI research), behavioural (how machines communicate in a deceptive manner), cognitive (how machines 'think' when they attempts deception), socio-ethical (how machines relate to others around them and viceversa), and ecosystemic, i.e. how external evolutionary pressures influence all the previous aspects [71].

Deceptive AI also has multiple ethical facets. Deception, by definition, clearly falls into the category of dishonest and potentially unethical behaviour which opposes the current emerging trend of ethical design in AI. However it can also be beneficial to society because its ethics depends on the aim of the AI agents and the context in which they are allowed to operate [63].

In order to be able to reap the benefits of prosocial deceptive AI and avoid the negative effects of developing more advanced deceptive technologies, the AI community must not just prescriptively aim for the ethical design of machines, but it should continuously reflect on what these technologies truly are and on the consequences that emerge due to their deployment in real-world contexts.

REFERENCES

1. Jonathan Albright. Welcome to the era of fake news. *Media and Communication*, 5(2): 87–89, 2017.
2. Tobias Becker, Andreas Agne, Peter R Lewis, Rami Bahsoon, Funmilade Faniyi, Lukas Esterle, Ariane Keller, Arjun Chandra, Alexander R Jensenius, and Stephan C Stilkerich. Epics: Engineering proprioception in computing systems. In *2012 IEEE 15th International Conference on Computational Science and Engineering*, pages 353–360. IEEE, 2012.
3. Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
4. Margaret A Boden. *AI: Its Nature and Future*. Oxford University Press, 2016.
5. Molly Bohannon. Lawyer used ChatGPT in Court—And cited fake cases. A judge is considering sanctions. *Forbes Magazine*, 2023.
6. Thomas Bolander and Mikkel Birkegaard Andersen. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.
7. Jason Borenstein and Ron Arkin. Robotic nudges: the ethics of engineering a more socially just human being. *Science and engineering ethics*, 22(1):31–46, 2016.
8. Cristiano Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2(2):113–119, 2000.
9. Tathagata Chakraborti and Subbarao Kambhampati. (when) can ai bots lie? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 53–59, 2019.
10. Raja Chatila and John C Havens. The IEEE global initiative on ethics of autonomous and intelligent systems. In *Robotics and well-being*, pages 11–16. Springer, 2019.
11. Mark Coeckelbergh. How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of icts in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology*, 20(2):71–85, 2018.
12. Jay A Conger. The necessary art of persuasion. *Harvard Business Review*, 76:84–97, 1998.
13. Michael T Cox. Metacognition in computation: A selected research review. *Artificial intelligence*, 169(2):104–141, 2005.
14. Daniel Dennett. Why creative intelligence is hard to find. *Behavioral and Brain Sciences*, 11(2):253–253, 1988.
15. Daniel C Dennett. The role of language in intelligence. *Sprache und Denken/Language and Thought*, page 42, 2013.
16. Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature, 2019.
17. Rino Falcone, Munindar Singh, and Yao-Hua Tan. *Trust in cyber-societies: integrating the human and artificial perspectives*, volume 2246. Springer Science & Business Media, 2001.
18. Xavier Ferrer, Tom van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80, 2021.
19. Brian J Fogg. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 225–232, 1998.
20. Alex Garland. *Ex machina*. Faber & Faber, 2015.
21. Alison Gopnik and Andrew Meltzoff. Imitation, cultural learning and the origins of "theory of mind". *Behavioral and Brain Sciences*, 16(3):521–523, 1993.
22. Gian Maria Greco and Luciano Floridi. The tragedy of the digital commons. *Ethics and Information Technology*, 6(2):73–81, 2004.
23. Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 2023.
24. Jürgen Habermas. *The theory of communicative action: Lifeworld and systems, a critique of functionalist reason*, volume 2. John Wiley

& Sons, 2015.
25. Charles L Hamblin. Mathematical models of dialogue 1. *Theoria*, 37(2):130–155, 1971.
26. Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.
27. Stevan Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1:43–54, 1991.
28. Stevan Harnad. Minds, machines and turing: The indistinguishability of indistinguishables. *The Turing test: the elusive standard of artificial intelligence*, pages 253–273, 2003.
29. Stevan Harnard. Turing testing and the game of life: Cognitive science is about designing lifelong performance capacity not short-term fooling. *Impact of Social Sciences Blog*, 2014.
30. HLEG, in AI. Ethics guidelines for trustworthy AI. *B-1049 Brussels*, 2019.
31. Alistair Isaac and Will Bridewell. *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press, 2017.
32. Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. Coercion and deception in persuasive technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, pages 38–49. CEUR-WS, 2018.
33. Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. Implementing argumentation-enabled empathic agents. In *Proceedings of the European Conference on Multi-Agent Systems*, pages 140–155. Springer, 2018.
34. Samuel Kounev, Peter Lewis, Kirstie L Bellman, Nelly Bencomo, Javier Camara, Ada Diaconescu, Lukas Esterle, Kurt Geihs, Holger Giese, Sebastian Götz, et al. The notion of self-aware computing. *Self-Aware Computing Systems*, pages 3–16, 2017.
35. Ralph Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51, 2011.
36. David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
37. Timothy R Levine. *Encyclopedia of deception*. Sage Publications, 2014.
38. Peter R Lewis. Self-aware computing systems: From psychology to engineering. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, pages 1044–1049. IEEE, 2017.
39. Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *arXiv preprint arXiv:2310.19775*, 2023.
40. Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. Characterising deception in ai: A survey. In Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney, editors, *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*, pages 3–16, Cham, 2021. Springer, Springer International Publishing. ISBN 978-3-030-91779-1.
41. Michael D Miller and Timothy R Levine. Persuasion. In *An integrated approach to communication theory and research*, pages 261–276. Routledge, 2019.
42. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
43. Kevin D Mitnick and William L Simon. *The Art of Deception: Controlling the Human Element of Security*. John Wiley & Sons, 2011.
44. Simone Natale et al. *Deceitful media: Artificial Intelligence and social life after the Turing Test*. Oxford University Press, USA, 2021.
45. Ulric Neisser. The roots of self-knowledge: Perceiving self, it, and thou a. *Annals of the New York Academy of Sciences*, 818(1):19–33, 1997.
46. Elinor Ostrom. *Governing the commons: The*

evolution of institutions for collective action. Cambridge university press, 1990.
47. Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. Lies, bullshit, and deception in agent-oriented programming languages. In *Proceedings of the 20th International TRUST Workshop @ IJCAI/AAMAS/ECAI/ICML*, pages 50–61, Stockholm, Sweden, 2018. CEUR Workshop Proceedings.
48. Gordon Pask. Conversation theory. *Applications in Education and Epistemology*, 1976.
49. David Pereira, Eugénio Oliveira, and Nelma Moreira. Modelling emotional bdi agents. In *Workshop on Formal Approaches to Multi-Agent Systems (FAMAS 2006), Riva Del Garda, Italy (August 2006)*, 2006.
50. Nancirose Piazza and Vahid Behzadan. A theory of mind approach as test-time mitigation against emergent adversarial communication. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2842–2844, 2023.
51. Jeremy Pitt. *Self-organising multi-agent systems: Algorithmic foundations of cyber-anarcho-socialism*. World Scientific, 2021.
52. Stephen Porter, Leanne ten Brinke, Alysha Baker, and Brendan Wallace. Would I lie to you? "leakage" in deceptive facial expressions relates to psychopathy and emotional intelligence. *Personality and Individual Differences*, 51(2):133–137, 2011.
53. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://blog.openai.com/better-language-models/, 2019.
54. Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477, 2019.
55. David Resnick. The ethics of science. *London: Rout*, 1998.
56. Michele Rocha, Heitor Henrique da Silva, Analúcia Schiaffino Morales, Stefan Sarkadi, and Alison R Panisson. Applying theory of mind to multi-agent systems: A systematic review. In *Brazilian Conference on Intelligent Systems*, pages 367–381. Springer, 2023.
57. Kevin Roose. Here come the fake videos, too. https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html, 2018.
58. Stefan Sarkadi. *Deception*. PhD thesis, King's College London, 2021.
59. Stefan Sarkadi. An arms race in theory-of-mind: Deception drives the emergence of higher-level theory-of-mind in agent societies. In *4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems ACSOS 2023*. IEEE Computer Society, 2023.
60. Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin D. Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, 2019.
61. Ştefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, and Iyad Rahwan. The evolution of deception. *Royal Society Open Science*, 8(9):201032, 2021.
62. Stefan Sarkadi, Ben Wright, Peta Masters, and Peter McBurney (Eds.). *DeceptiveAI*, volume 1296. Springer, 2021.
63. Stefan Sarkadi, Peidong Mei, and Edmond Awad. Should my agent lie for me? a study on attitudes of us-based participants towards deceptive ai in selected future-of-work scenarios. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2023.
64. Jordan Richard Schoenherr. The first total war and the sociotechnical systems of warfare. *IEEE Technology and Society Magazine*, 42(3):42–56, 2023.
65. Elizabeth Sklar, Simon Parsons, and Mathew Davies. When is it okay to lie? a simple model of contradiction in agent-based dialogues. In *ArgMAS*, pages 251–261. Springer, 2004.
66. Phillip Staines. *Linguistics and the Parts of the Mind: Or how to Build a Machine Worth Talking to*. Cambridge Scholars Publishing,

2018.
67. Rod Thornton and Marina Miron. Towards the 'third revolution in military affairs' the Russian military's use of AI-enabled cyber warfare. *The RUSI Journal*, 165(3):12–21, 2020.
68. Alan Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, 1950. URL http://www.jstor.org/stable/2251299.
69. Henry M Wellman, David Cross, and Julanne Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.
70. Eliezer Yudkowsky. The AI-box experiment. *Singularity Institute*, 2002.
71. Xiao Zhan, Yifan Xu, and Stefan Sarkadi. Deceptive ai ecosystems: The case of chatgpt. In *Conversational User Interfaces, CUI'23, July 19–21, 2023, Eindhoven, Netherlands*. ACM Proceedings, 2023.

Ștefan Sarkadi is currently a Proleptic Lecturer and Research Fellow at King's College London, UK. He has a multidisciplinary background in Philosophy (B.A.), Cognitive Science (MSc.), and Computer Science (PhD). Ștefan's PhD thesis is the first full computational treatment of deception at the intersection of Artificial Intelligence, Philosophy, and Psychology, and he has been doing research on the topic of Deceptive AI since 2016. His research on deception covers areas such as intelligence analysis, multi-agent systems, agent-based modelling, cognitive AI architectures, argumentation, and human-AI interaction. Ștefan is a member of the Royal Academy of Engineering through the RAEng UK Intelligence Community Research Fellowship. Contact him at stefan.sarkadi@kcl.ac.uk.