Trust in AI: A Unified Approach

Andreas Kaminski

Abstract *The discourse on trust in technology, and especially in AI, has crystallized around two dominant positions. According to the first, technology or AI cannot be trusted and doing so would constitute a category mistake. In response, a second position has emerged, which interprets trust in technology as reliability. This represents, in effect, a retreat: it responds to the critique of the first position by adopting a significantly narrower epistemic concept of trust. This paper acknowledges the initial critique (i.e., the category mistake) as valid. However, it argues that retreating to a reliabilist understanding of trust is not the only available option. An alternative remains open, one that allows for a richer concept of trust, grounded in virtue theory. This alternative shifts the focus to the relationship between individuals, organizations, and technology. The paper outlines the basic contours of this alternative approach to understanding trust in AI.*

Trustworthiness of AI as well as trust in AI has become a title of numerous publications (AI HLEG 2019; Floridi 2019; Thiebes, Lins and Sunyaev 2021). The title presupposes that talking about trust in technology makes sense which is the subject of ongoing debates (Wagner 1994; Nickel, Franssen and Kroes 2010; Kaminski 2010; Floridi 2019; Thiebes, Lins and Sunyaev 2021). This has been doubted, and for good reasons (cf. Lahno 2002: 134–138). Some theories of trust have presented cogent arguments that this way of speaking constitutes a category mistake; that is: applying the concept of trust to technology in general and AI in particular is fundamentally misguided. According to these arguments, speaking of trust in technology is as nonsensical as attributing diabetes to a number or declaring a rock on the moon bankrupt. At best, we could try to read such statements metaphorically.

Given these arguments, it seems that any defense of trust in technology must rely on a substantially deflated or impoverished version of the concept of trust. A natural course, then, appears to be limiting the concept to a specific epistemic variant. Reliabilist approaches in particular appear to offer a viable path for reconciling the critique with a coherent way of talking about trust (Durán and Formanek 2018). On this understanding, a system is reliable when, in most cases, it produces the desired effect or serves the intended purpose. Reliabilist approaches therefore con-

strue the trustworthiness of systems primarily as something that can be recognized and even measured – above all by their track record. This comes with the aforementioned limitation: Trust is understood in a different, restricted sense within these epistemic theories. The limitations include that the specifically normative dimension of trust and the special form of relationship that exists between persons who trust each other cannot be captured. When trust is disappointed, the response is mainly (or at least also) morally determined. The epistemic approach, however, cannot make sense of this. On that view, whether we trust or not depends solely on our epistemic expectations.

In trust research, it has become something of a standard move to treat trust as a multifaceted social phenomenon, requiring different theoretical accounts depending on the context – whether between friends, market actors, or in relation to God or technology. From this standpoint, it seems reasonable to propose a distinct concept of trust that, while not capturing all phenomena, at least makes sense of talking about technology.[1]

In the following, I want to show three things:

1) The arguments that a certain way of speaking about trust in technology constitutes a category mistake are compelling;
2) Nevertheless, there is a different understanding of what it means to say that we trust technology – one that does not amount to a category mistake;
3) Moreover, this way of speaking avoids a reductive understanding of trust but maintains its normative significance.

I begin by outlining the basic features of epistemic trust theories. Subsequently, I address the criticism of epistemic trust theories, which simultaneously forms the basis for the arguments why trust in technology is not possible. Then I elaborate on the fallback position that trust in technology should therefore be understood merely epistemically. The following section will then present the alternative position. Finally, I outline what this means for responsibility and the capacity-based approach.

1 This routine was able to spread easily within the humanities and social sciences since theoretical pluralism is taken for granted in these disciplines. Moreover, there are specific reasons tied to the theoretical history of trust. At the beginning of the current wave of intense engagement with trust, Annette Baier raised a criticism in an essay that, in many ways, prefigured later debates: Namely, that when trust had been a subject of philosophical reflection at all, it had typically been understood through the model of a contract (Baier 1986). That is, as a situation conceived symmetrically, where two equal parties enter into an explicit agreement. This implicit orientation toward the contract model, however, is ill-suited to many other forms of trust – for instance, those between parents and children. A plurality of forms of trust, it was argued, should be matched by a plurality of theories of trust (cf. Frevert 2013: 9; Hardin 2002: 6; Hartmann 2001: 7 et seq.).

Epistemic Theories of Trust

Epistemic theories of trust come in various versions: the so-called evidential view (examples are Hume 2007; Coady 1992) or rational choice and game theory (Coleman 1990; Gambetta 1988; Hardin 2006). What they have in common is that they conceptualize trust as a relationship between A and B in which A *knows* something about B that leads A to expect that B will act in a way that benefits A.

In the empirical version of the evidential view, for instance, for David Hume (Hume 2007, ch. X) or in the psychology of testimonial assertions (Stern 1903; Sporer 1997; Arntzen 2007), this knowledge on A's part might consist of B's track record of acting in a trustworthy manner. It could also be based on indicators that show whether B will act in a trustworthy way. A prime example would be credit models (Lauer 2017). These provide A with a track record, derived from experience, showing how often B has taken out loans and repaid them. A can also use indicators to classify B into a (statistical) group for which there is a known track record; for instance, B's place of residence, level of education, age, number of children, or the presence of a car (kept in a garage). According to this approach, entirely different indicators could also be used, such as those analyzing B's facial expressions as signs of truthfulness or a deceptive intent in communication (Ekman 2001, 2003; Kaminski 2020b).

According to this theory, trust is based on a foundation that is *independent* of trust itself. This foundation is the evidence for B's trustworthiness or lack thereof: A *knows* whether B is trustworthy. Trust is then the inference drawn from this evidential basis: a forecast giving rise to a cognitive expectation. When conceptualized in this way, trust becomes indistinguishable from mere prediction (of a positive outcome). Such a concept of trust or a method of establishing trustworthiness can be applied to persons, institutions, plants, animals, technology, etc. The underlying reason is that, essentially, we are dealing only with *inductive inferences*[2] that are understood from a *practical perspective*, insofar as they guide decisions. Trust is the natural conclusion when the expectation is that a positive outcome will occur; conversely, if a positive outcome is not anticipated, the result is distrust. From this, we can already provide a characterization of what trust means in this epistemic version: *Relying on the other because one relies on the procedure.* Primarily, it is *not* the other person (or the technology depending on what occupies the other position in the relationship) that is trusted, but the evidence itself and the process of drawing correct conclusions from it. One relies on B by, for example, relying on the reliability of the indicators. The counterpart's reliability depends on the reliability of the (inductive) procedure. Naturally, different indicators can have varying degrees of reliability.

2 This was already clear to Hume. See his remark in the *Enquiry*, where he notes that these inferences are the same as those drawn when discovering the connection between cause and effect (Hume 2007: 80).

Cognitive expectation, cognitive disappointment: Because it is strictly a cognitive expectation, disappointment is similarly cognitive in nature. One might just be surprised or respond with understanding if the expectation was never very certain in the first place, since disappointment can be expected in such cases (at roulette, sometimes the ball lands on red, sometimes on black). Within this framework, one can never be *morally* disappointed.[3]

More evidence, more trust: If a positive outcome can be expected, more evidence leads to greater trust. It does not matter what the evidence is based on or how it is obtained: Threats of sanctions or surveillance could, in principle, boost trust just as much as the other person's virtues.

Other variations of epistemic theories, such as rational choice or game theory, remain within the same framework. The main characteristics are the same. Additional assumptions may appear, for example (i) regarding the other person's interests, (ii) that the other person will pursue these interests rationally, and (iii) in game theory in particular, calculations that compare one's own expectations with those of others. Nevertheless, these additions do not alter the epistemic framework; they simply refine it. Thus, Coleman can regard the question of whether one should trust others as a special case of decision-making under risk, conceptualized in the form of a *bet* (Coleman 1990: 99). For rational choice and game theorists, the trustworthiness of the other person ultimately reduces to a probability value (Dasgupta 1988; Gambetta 1988).

3 The reason for this is not that one is dealing, for instance, with economic relations, but rather that the theory itself lacks a framework for articulating a moral violation. If an evidence-based prediction is made and the claim contained in the prediction does not come true, the person who made the prediction cannot be morally disappointed that the object whose behavior was predicted did not act as expected. The reference to the object (an evidence-based prediction) does not allow for the formulation of a moral disappointment. This holds regardless of whether the prediction concerns the course of the stars or the behavior of people. Only when the reference is no longer purely epistemic can the possibility arise for expressing moral disappointment. Now, consider a case in which a person enters into an economic cooperation based on the game-theoretic prediction that their potential partner will act cooperatively – and is then disappointed. In this case, game theory does not provide grounds for moral disappointment. At most, the disappointment could serve as a kind of sanction, which in turn constitutes an epistemic reason; namely, a factor intended to promote current or future cooperative behavior. This can be taken into account in future predictions, in the sense that the threat of sanction is treated as a motivating factor that increases the likelihood that the other person will act cooperatively. To be sure, people who reason in game-theoretic terms can in fact experience moral disappointment. But they can not do so on the basis of game theory. Their moral disappointment rather reveals that they relate to the other person in a way that is not purely epistemic.

Critique of Epistemic Theories of Trust

Within epistemic theories, the role of evidence and the resulting forecasts leads to a conception of trust that does not fundamentally differ from predictions about the behavior of stars. One 'trusts' that a certain star will appear in the same place tomorrow, just as one trusts that a friend will show up in time to pick one up to go to the airport. Ultimately, as we have seen, trust is not placed in the other person but in one's own cognitive faculties, which have carefully evaluated the evidence. This basic situation does not change even if we factor in second-order expectations as in game theory. Such considerations merely complicate inference but do not alter the nature of the relationship between A and B.

The critique of epistemic trust theory, advanced, for instance, by the so-called *assurance view* or by theories that treat trust as a feeling, therefore focuses on the nature of the relationship. I highlight three criticisms, which form a family of objections insofar as they all center on this relational aspect:

1. *Not actually trusting the other person:* In epistemic theory, trust in a car or in a friend can be assessed in essentially the same way: Both could, for example, have a good track record. The car always starts, and the friend always shows up when needed. The reason is that, within epistemic theory, the other person ultimately is not considered *as a person*. Indeed, the more thoroughly one understands the rules that govern another's actions, the better one can assess their trustworthiness. These rules gain in evidential force the more they resemble identifiable mechanisms (Lahno 2002: 168–170). However, once another's actions are interpreted merely as the outputs of an underlying mechanism, the person is no longer seen as an autonomous agent but rather as analogous to a natural process or technical system. In contrast, non-epistemic accounts emphasize the autonomy of the other as the foundation of their trustworthiness (Lahno 2002: 169–178; Moran 2006: 278 et seq.).

2. *The normativity of trusting relationships:* If trust ultimately rests on the reliability of a mechanism (producing a merely cognitive expectation), then its distinct normativity cannot be adequately captured. A person whose trust has been betrayed is not simply surprised in a factual sense but wronged in a normative one. The disappointment is not just interpreted as the truster's mistake, e.g., an overestimation of the available evidence – that is possible too – but at least also as a failure on the part of the trusted person. What is at stake is not merely the expectation that p will occur, but the expectation that you will ensure that p. I do not merely expect *that* p, I expect *from you* that p (Faulkner 2007: 881). In this light, disappointment becomes a breach of trust, a "betrayal" (Lagerspetz 1998: 42). Trust presupposes that the other person assumes responsibility voluntarily and autonomously within the relationship (Moran 2006: 281). Only what someone takes on intentionally and freely can genuinely be the object of trust.

3. *Impermissible epistemic reasons:* Because epistemic theories ground trust in epistemic reasons, certain consequences arise that run counter to the very nature of trust. Increasing sanctions or threats of sanctions can strengthen these epistemic reasons, which, within the epistemic perspective, ought to enhance trust. Likewise, from an epistemic point of view, surveillance would justify "more trust", insofar as it provides better informational grounds.[4] The same essentially applies to increasing control over the other person. Someone who reads another person's diary, monitors them at every turn, and threatens them with severe penalties will, from an epistemic standpoint, have a better basis for their 'trust' (Kaminski 2020a: 181; 2023). Clearly, if trust is supposed to rest on a voluntary assumption of responsibility, such epistemic grounds are excluded.

The fundamental error in all three cases can be traced back to the same category mistake: Epistemic theories use categories suited to causal prediction of natural or technical processes to describe interpersonal relations grounded in autonomy. They base their understanding of trust on procedures that do not treat others as autonomous persons but rather assess or measure their 'trustworthiness' in light of their predictability analogous to mechanisms. Trust, by contrast, requires understanding others as autonomous persons. If the freedom of the other is not placed at the center, predictions – especially those grounded in technomorphic forms of control – are easily mistaken for trust. This does not mean that others cannot be reliable. One can form expectations, but these rest on fundamentally different grounds: not on mechanical patterns of behavior, but on the other person's autonomous decision to be trustworthy.

The 'Reliability Residuum' and Its Problems

In light of the aforementioned critique, it may seem that there is no longer a meaningful way to speak of trust in technology and thus in AI systems. The criticism of the epistemic theory of trust showed that the form of the relationship between two individuals who trust each other cannot be replicated by technical instruments, machines, or systems. For such replication to occur, technology would have to be an autonomous subject capable of recognizing and voluntarily entering into normative

4 More precisely put: This applies in cases where surveillance reveals that the other person intends to cooperate, harbors no harmful intentions, and so on. However, the central point remains unaffected: Trust and direct knowledge are mutually exclusive. If I watch the other person and *see* that they are attending to my concerns, I no longer need to trust them. The same holds for control: Trust and control are mutually exclusive. If I control the other person's behavior – or even merely attempt to do so (for instance, by threatening sanctions) – then I am not trusting them. From within the epistemic approaches, however, there is no clear reason why such surveillance or control would not serve to justify (or even promote) trust.

commitments. Someone who trusts a given technology would thus not merely base their trust on knowledge about that technology; they would base it on the relationship they have with the technology. In that relationship, they would not only expect a certain effect to occur when initiating a process, but also expect the technology to care that this effect occurs – precisely because it is trusted.

It is obvious that technology does not and cannot fulfill this condition. Otherwise, one would have to assume that technology enjoys the status of an autonomous subject that does not merely operate according to rules we can discern but also can place itself in a normative relationship by acknowledging normative rules and acting in accordance with them. Situations in which technology is sometimes assumed to meet this condition appear to arise from an ambiguity – namely, equating the moral autonomy of persons with the technical autonomy of systems, whose functioning may not be determined directly by developers but is governed by higher-level rule systems that generate the operative rules only afterward (Kaminski 2014).[5] However, this question cannot be pursued further here, so I will proceed on the assumption that technical systems do not exhibit moral autonomy.

If this is the case, then only a weaker sense remains in which one could speak of 'trust' in technology. And that sense seems to be delineated precisely by the epistemic theory. Since the latter does not capture trust in a substantive sense, a restricted concept must take its place. Hence, the discourse has turned to the reliability of technical systems. On the one hand, reliability was introduced, albeit in an unclear manner, by Annette Baier as the opposite of trust in a narrower sense at the start of the modern debate (Baier 1986). On the other hand, reliability is also a familiar technical concept. For instance, AI models have long been tested to see how reliably they carry out specific classifications like identifying objects in images; currently large language models are assessed in terms of their reliability in generating accurate responses.

Here we rediscover Hume's idea of a track record: The degree of trustworthiness we ascribe to someone who promises something is evaluated by the ratio of total promises made to promises actually fulfilled. Analogously, the degree of reliability of image recognition is determined by comparing the total number of image-recognition processes with the number of correct recognitions.[6]

Should we therefore confine our talk about the trustworthiness of technology to this notion of reliability? Must we do so, given the arguments that rule out any

5 In the case of machine learning algorithms, the base architecture, hyperparameters (such as the learning rate), cost function, and data are specified by developers but not the learned model itself; that is the result of the training process. As a consequence, what emerges is a kind of program that was not directly coded (Angius and Plebe 2023).

6 In AI literature, the terminology may differ insofar as there is also talk of "robust accuracy". However, this is merely a terminological difference, not a conceptual one.

broader understanding? One initial problem is that many debates on the *trustworthiness* of AI cannot be limited to whether the technology reliably detects and classifies objects in data. While this is *one* important dimension, it is by no means the only relevant factor for deciding whether a model is trustworthy. An AI system that consistently reaches unjust decisions would hardly be considered trustworthy; likewise, a chatbot that reliably deceives users would not merit that label – at least not by those who consider justice or veracity important. A system that is indeed reliable and just but fails to protect sensitive data would likewise not be deemed trustworthy.

Hence, we encounter two claims that appear to clash:

1. We can talk about technology as reliable, but there is no meaningful sense in which we can call it trustworthy.

2. Reliability is not sufficient for how we assess the trustworthiness of technology in our discussions.

One could attempt to resolve this terminologically. But that would be unlikely to succeed, for, as point (2) shows, there is a significant way of talking about technology that does not limit itself to reliability. Faced with this conceptual dilemma, we initially have two options:

a. We can either try to reduce all aspects of trustworthiness (2) to reliability (1).
b. Or we can take the opposite approach and differentiate the layers, attributing (1) to the technical sphere and (2) to the social sphere.

However, neither option seems promising.

Let us start with the idea of *differentiating* levels: The solution would be to say that we attribute reliability to the technical system while ascribing broader trustworthiness – such as meeting standards of justice or protecting sensitive data – within the social dimension; presumably, this would mean that we trust the developers or providers of these systems. This may not be entirely wrong, but it oversimplifies the fact that we do, in fact, trust the system itself to make fair decisions or adequately protect data. After all, it is not the developers themselves who make these decisions in real time; the system must provide the security measures that prevent unauthorized access to the data it contains.

The alternative solution would be a *reduction*: Justice or data privacy (or privacy more generally) would be regarded as special cases of system reliability. One could flesh out this idea as follows: A system is fair if it reliably produces fair decisions; it protects data if it reliably ensures that only authorized data accesses occur. In this way, reliability would become the overarching framework for evaluating systems (and people) in numerous dimensions. What we previously labeled reliability – the system's functional correctness, for instance – would itself be just one special case (which one might call functional reliability or robust accuracy). If the function

of the system is to classify objects in image data, one can then say the system is reliable to such-and-such a degree at fulfilling that function.

A central difficulty with this approach, however, is that interpreting justice or data protection as mere reliability can gloss over serious violations by focusing only on the quantity of failures rather than their quality. A system that is "nearly always" fair or that "nearly always" protects data but occasionally allows severe transgressions would, by this definition, still be a reliable system – even if the exceptions were catastrophic. One might try to compensate for this by taking the severity of individual cases into account, but then one would be leaving behind the tightly circumscribed lines of a reliabilist approach, since it would require engaging in substantive debates about, for example, the nature of justice. Such an attempt at stabilization would naturally lead to what I am about to present.

Practices and Outcomes in the Light of Virtues

How can we resolve the problem of trust in technology, given the dead ends reached by the earlier proposals? We have seen that talk of trust in a substantive sense goes beyond mere reliability and that it seemingly applies only to autonomous persons – those who understand that they are trusted, who voluntarily enter into trusting relationships and who can acknowledge the normative commitments and expectations these relationships entail. Yet the same observation appears to hold true for the predicate 'just'. Much like the term 'trustworthy' seems to presuppose personhood, one might assume that 'just' is likewise meaningful only when applied to persons. But this is not in fact the case: We speak of just rules without thereby granting the rules themselves a subject status in this sense. Should we, by the same logic, restrict our talk of 'just' in the same manner suggested by the above argument concerning 'trustworthy'?

My thesis is that precisely the opposite is true. We can learn something about the trustworthiness of technical systems from the way we say that certain rule systems are just or unjust. When we call a system of rules just, we mean more than that it reliably produces just decisions (although that is a necessary condition, as established in Section 3, it is not sufficient). Nor do we ascribe subject status to them. Rather, I suggest that what we mean is that these rules result from deliberations conducted in the spirit of justice – deliberation about, say, the fair distribution of resources. The rules were developed with the virtue of justice in mind. They are not accidentally just; rather, they were deliberately shaped to embody a 'just spirit'.

Another way to understand this proposal is to see it as drawing on the idea of a *pros hen*-structure.[7] In other words, there is a focal meaning of discourse about virtues: Virtues primarily concern a person's character. Yet we can also say of a person's actions that they are just or unjust and of the consequences of those actions that they result in just or unjust outcomes. By the same token, we can speak of AI systems and technology in general as manifesting certain virtues.

This proposal will (and should) prompt a series of follow-up questions. These concern the relationship between persons and technical systems, the connection between trustworthiness and other virtues, the epistemic and normative dimensions of trustworthiness, and finally, how one might establish whether something is in fact virtuous.

Let us begin with the question of how persons and technical systems relate. In most cases, we do not know the developers of technical systems personally. How, then, can we say anything about the systems' virtue as derived from their developers? Put differently, if we describe a technical system as trustworthy while grounding that concept (via a *pros hen*-structure) in the notion of a person's character, yet lack any knowledge of the persons who developed or deployed the technical systems, how meaningful can it be to speak of 'trustworthy systems' at all? Here again, thinking about just rules offers a helpful analogy. To judge the justice of rules, we need not know the people who developed them. We can analyze the rules themselves and conclude whether they are just or unjust. An example would be gerrymandered voting districts – drawn so that one group's votes count disproportionately. Determining whether voting districts are unjustly structured does not require knowledge of who drew them. Such knowledge can, however, serve two purposes: First, it might prompt us to investigate further, if we know these particular individuals have engaged in such practices before (or used other unfair campaign tactics). Second, it may be needed to understand whether the injustice was a lapse or mistake, the result of incompetence, or a malicious, intentional attempt to make certain votes 'count more'. The same holds for AI systems. Whether a system produces fair decisions ultimately must be shown by examining those decisions themselves. Deciding whether an unfair system was simply sloppily developed or was intentionally designed to be that way may indeed involve knowing more about the developers – though in some cases, the system's very structure might reveal enough on its own.

7 Aristotle develops this idea about the structure of central concepts in our language in his *Metaphysics* with reference to the manifold ways in which we speak of being (cf. Aristoteles 1994: IV.2, 1003a2, 1003a33–1003b19). *Pros hen* could be translated as "in relation to one". For instance, there are various meanings of the term 'healthy' or 'health'. We speak of healthy food and of a person being healthy. In the former case, "healthy" refers to a cause: food that promotes health while in the latter, it denotes a result or state. According to Aristotle, these meanings are not unrelated. Rather, they can be organized with reference to a focal meaning: the healthy body. It is this central case from which the other uses derive their meaning.

The core idea, then, is this: We can trust technical systems when we view them as having been developed in the spirit of trustworthiness. That is, they were created with the explicit aim of making them trustworthy. In adopting this perspective, we are shifting our frame of reference: In evaluating trustworthiness, what we truly assess is whether the system's development was guided by the values that should be central to creating trustworthy technologies. The systems are the result of practices directed by developers who wanted to produce trustworthy systems. The analogy to rules that emerged from processes oriented toward the virtue of justice is illuminating here as well. However, acknowledging the significance of development processes (and thus the developers themselves) does not imply that people can trust only those systems whose developers they personally know. The very design of the systems can bear the marks of orientation toward such virtues, and this is what our trust rests on.

Imagine someone who is analyzing a technical system and finds that it was designed in such a way as to make it particularly difficult to hack. In such a case, it is of course appropriate to say that the system was designed with a focused concern for the value of security – that it was developed in that spirit. Someone who speaks this way need not be presumed to have any personal acquaintance with the system's developers.

This raises the question of who is in a position to analyze a system in such a way as to determine which values it embodies and which values, therefore, informed its design. When I look at a chair, I might still have a few cues that help me judge whether it was designed to be robust, comfortable, or whether a low price point was the primary consideration. Many other aspects, such as ecological values, tend to elude direct assessment. This is all the more true for information technology, and especially for complex digital systems: Our capacity for evaluative judgment becomes markedly more constrained when we turn to AI models. To evaluate the values according to which such systems are designed, we are far more dependent on the expertise of others. One can nonetheless see the general direction that an assessment of trustworthiness might take. However, within the context of this article, the question is not how such an assessment could be carried out in practice. Rather, the focus is on the theoretical question: Is there a meaningful way to speak of trust in technology in general, and trust in AI in particular? I believe I have now sketched the essential contours of the answer. The task now is to develop this outline in more detail.

Trustworthiness as the Unity of Other Virtues

What kind of virtue is trustworthiness? When we consider trustworthiness alongside other values such as justice, privacy, and reliability, each value appears to address a specific aspect of the system: Reliability measures how consistently a sys-

tem performs its intended function, privacy concerns how well data is protected against unauthorized access; justice assesses whether decisions, for example, about the allocation of resources, are made in a way that ensures no individual is disadvantaged.[8] Is there a particular aspect devoted to the trustworthiness of systems, alongside other aspects such as justice or privacy? Initially, it might look as though this aspect is empty, as though it cannot be filled with any concrete content. But this perception stems from the assumption that trustworthiness is just one additional segment – expressed in requirements-engineering terms, one non-functional requirement among others.

The confusion disappears if we look at how breaches of trust are typically discussed: "M was not honest", "N did not have the courage to stand up for me", "O did not care that their decision was unjust". Such statements might explain why someone lost trust in another. If we consider these explanations closely, we see that trust relates to other values. For instance, when we trust someone, we often trust that a friend will be *honest*, that a colleague will be *sincere*, that a partner will *courageously* stand up for us, that a doctor will be *careful*, that a supervisor will be *fair*, and so on. From this follows the idea that trust does not stand alongside other values or virtues but is instead linked with them in the following way: *The object of trust is the other person's trustworthiness, and trustworthiness in turn relates to other values or virtues.* A trustworthy person will prove to be honest, just, courageous, and kind when it matters. In short, trustworthiness is the unity of other virtues; it is not just one virtue among many, but rather the social connection point at which they come together (Kaminski 2020a: 396 et seq.)[9]

By analogy, the trustworthiness of technical systems is not merely one non-functional requirement among others such as privacy; rather, it is their overall unity. When we trust technical systems, we trust them to be trustworthy in all the relevant areas; the relevant areas are precisely those values we consider essential – such as ensuring no discriminatory bias, avoiding unjust recommendations, protecting data, reliably performing their function, and so forth. From this perspective, it is hardly surprising that lists describing "trustworthy AI" often lack a clear systematization of the values at stake. Trustworthiness is oriented toward the specific values and virtues that matter in each particular context. Which ones matter depends on

8 See, for example, the lists of values such as the requirement that technical systems be "effective, interoperable, robust, and reliable" (Art. 50, No. 2, AI Act), or that they ensure "accuracy, reliability, and transparency" in order "to avoid adverse impacts, retain public trust" (Recital 59, AI Act).

9 This raises the question of how this thesis relates to Aristotle's idea of justice as a unity. If we understand justice as the universal appropriateness of one's actions, then this apparent conflict may be smaller than it now seems (cf. Aristoteles 2011: 1129a).

the situation, something that, as Aristotle remarks, can be specified only in general outlines (Aristoteles 2011: 1104a).

Even though the formal object of trust is thus the trustworthiness of persons or technical systems, and that trustworthiness points toward other values or virtues, this does not automatically mean that trust must always be understood as a three-place relation: A trusts B to do p. Granted, a form of trust that never manifests in any concrete relation (p) would not only be unrecognizable but would eventually cease to exist. I see the relationship between the three-place and two-place formulations of trust as a modal relation (actual vs. possible). Trust does manifest itself concretely in particular instances, but it is not reducible to them since the trust relationship creates (and also delimits) the possibility of various future scenarios. And precisely because situations can only be anticipated in broad outlines, it is impossible to specify exhaustively in a finite list all the things our trust might someday encompass.

The Epistemic and Normative Dimension of Trust

The notion of trust in technology introduced here, which, from my perspective, no longer entails a category mistake, goes hand in hand with a different theory of trust. Besides the epistemic theory of trust (in its various forms), there is a well-established normative view of trust. This normative view appears in different guises: the assurance view (Moran 2005, 2006, 2013; Ross 1986) or as an affect- or emotion-based theory of trust (Lahno 2002; Faulkner 2015). The normative theory represents a profound transformation in how trust is understood, making substantial revisions to the epistemic perspective by avoiding its abstractions. However, normative accounts suffer from other one-sidedness. Their greatest weakness is that, in most of their versions, they fail to place the epistemic dimension of trust in a coherent relationship with its normative dimension (Kaminski 2017; Wiengarn 2021).

Yet the phenomenon of trust itself demands a conceptual framework that captures the unity of these dimensions. A crucial indication of this emerges from the experience of disappointed trust. When our trust is disappointed, it evokes two responses: (1) How could I have been so mistaken about you? and (2) How could you have deceived me like that? In short, we respond by questioning our own judgment (What could I have done better? How can I learn to grant trust more judiciously?) and, at the same time, by blaming the other person (You shouldn't have done that!). Here, reflection and learning on the one hand, and the moral gravity of betrayal on the other, converge in the way we may respond. This raises the question of what a

theory of trust capable of grasping this unity might look like. Normative theories can only account for the normative response. They miss half of the phenomenon.[10]

The contours of such an approach, one that unifies our responses and captures trust in a non-reductive way, have already been implicitly drawn in references to the virtues of trustworthiness and trust. This is a third, alternative perspective (Kaminski 2020a, 2023), one that does not adopt a dualistic view of trust from the outset. Ascribing a virtue to someone presupposes that we have observed their behavior in ways that justify attributing that virtue to them. We rely on epistemic reasons grounded in how they act. Yet these are only epistemic reasons for attributing a virtue if the person in question does not stand in a merely external relationship to that virtue – say, behaving 'as if' being virtuous by chance or for strategic purposes alone. Instead, the person must incorporate a normative relationship to the virtue in question and recognize it as a virtue. The same is true of the person who ascribes a virtue: In order to ascribe a virtue rather than merely calling someone or something a reliable mechanism of future behavior, that person must also recognize the relevant trait as a virtue, thereby adopting a normative stance toward it.

Virtues and the Capacity Approach

Within the virtue-ethical framework, the epistemic dimension of trust finds its proper place. This resolves the important theoretical challenge of how to understand both dimensions of trust in a unified way. It does not, however, solve the practical challenge of determining whether a person – or, in our context, a technical system – is trustworthy. This challenge is especially daunting because it is difficult to distinguish between behavior that appears virtuous but is actually driven by strategic motives (for example, someone who is friendly only so long as the boss is present) and behavior that is genuinely motivated by an orientation toward virtue. Typically, familiarity can make this easier: You notice, for instance, that as soon as the boss leaves the room, your colleague's recently displayed friendliness evaporates.

10 At times, the literature tends to conflate distinct aspects. Alvarado (2023), for instance, appears to conflate the object of trust (information) with the type of trust (epistemic trust) when he argues that, since the primary function of AI is to serve as an epistemic enhancer, trust in AI is therefore epistemic in nature. But even if the latter were true – i.e., even if the sole function of AI were to provide information (which it is not, given that such systems are also used in decision-making contexts) – it would by no means follow that the manner in which AI is trusted must be exclusively epistemic. The sincerity of another person, for instance, is not trusted merely in an epistemic sense; the same holds for promises concerning the reliability of an AI system – they, too, need not be trusted solely on epistemic grounds.

Ordinarily, becoming familiar with a technology is one of the central aims of its design and a natural outcome of practical use (Kaminski 2010: Part II). However, the rapid pace of technological development, particularly in AI, together with the opacity of models and massive marketing efforts, significantly complicates this process. If we follow the argument set out above, the capacity approach would need to focus on examining how and to what extent the virtues relevant to the specific model in its context (justice, privacy, reliability, and so on) are realized by these systems, and also whether the design and operation of these systems are in fact guided by an orientation toward those virtues.

Clearly, this task overwhelms individual users. Often, there is not even a theory-independent way to explain what fair decisions by AI models might involve. Consequently, the capacity approach can only be understood as an institutional and collective endeavor. In turn, a well-grounded decision about whether to trust an AI system presupposes that the institutions, media outlets, and communities evaluating that system are themselves worthy of trust. Each individual's capacities are bound up with the capacities of the broader network. Hence, justified trust in AI systems requires justified trust in these networks. For this reason, AI ethics ultimately is AI politics. Hence trustworthy AI depends on digital governance in a broader social network.

References

AI HLEG (2019): Ethics guidelines for trustworthy AI. https://data.europa.eu/doi/10.2759/346720, last access: July 25, 2025.

Alvarado, Ramón (2023): "What Kind of Trust Does AI Deserve, If Any?", in: AI and Ethics 3(4), pp. 1169–1183.

Angius, Nicola and Plebe, Alessio (2023): "From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning", in: Philosophy & Technology, 36(3), pp. 36–47.

Aristoteles (1994): Metaphysik, Neuausg. Reinbek b.H.: Rowohlt.

Aristoteles (2011): Nikomachische Ethik, 3rd ed., Wolf, Ursula (ed.), Reinbek b.H.: Rowohlt.

Arntzen, Friedrich (2007): Psychologie der Zeugenaussage. System der Glaubhaftigkeitsmerkmale, München: Beck.

Baier, Anette C. (1986): "Trust and Antitrust", in: Ethics 96(2), pp. 231–260.

Coady, Cecil (1992): Testimony. A philosophical study. Oxford: Oxford University Press.

Coleman, James S. (1990): Foundations of Social Theory, 1st ed., Cambridge, MA: Harvard University Press.

Dasgupta, Partha (1988): "Trust as a Commodity", in: Diego Gambetta (ed.), Trust. Making and Breaking Cooperative Relations, New York, NY: Basil Blackwell, pp. 49–72.

Durán, Juan M. and Formanek, Nico (2018): "Grounds for Trust. Essential Epistemic Opacity and Computational Reliabilism", in: Minds and Machines 28(4), pp. 645–666.

Ekman, Paul (2001): Telling Lies. Clues to Deceit in the Marketplace, Politics, and Marriage, New York: Norton & Company.

Ekman, Paul (2003): Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional life, New York, NY: Holt & Company.

Ekman, Paul and Rosenberg, Erika L. (eds.) (2005): What the Face Reveals. Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford: Oxford University Press.

Faulkner, Paul (2007): "On Telling and Trusting", in: Mind 116(464), pp. 875–902.

Faulkner, Paul (2015): Knowledge on Trust, Oxford: Oxford University Press.

Floridi, Luciano (2019): "Establishing the Rules for Building Trustworthy AI", in: Nature Machine Intelligence 1(6), pp. 261–262.

Frevert, Ute (2013): Vertrauensfragen. Eine Obsession der Moderne. München: Beck.

Gambetta, Diego (ed.) (1988): "Can We Trust Trust?", in: Id. (ed.): Trust. Making and Breaking Cooperative Relations, New York, NY: Basil Blackwell, pp. 213–237.

Hardin, Russell (2002): Trust and Trustworthiness. New York: Russell Sage Foundation.

Hardin, Russell (2006): Trust. Cambridge: Polity.

Hartmann, Martin (2001): "Einleitung", in: Hartmann, Martin and Offe, Claus (eds.), Vertrauen. Die Grundlage des sozialen Zusammenhalts. Frankfurt/Main: Campus, pp. 7–34.

Hume, David. (2007): An Enquiry Concerning Human Understanding, Oxford: Oxford University Press.

Kaminski, Andreas (2010): Technik als Erwartung. Grundzüge einer allgemeinen Technikphilosophie, Bielefeld: Transcript.

Kaminski, Andreas (2014): "Lernende Maschinen: naturalisiert, transklassisch, nichttrivial? Ein Analysemodell ihrer informellen Wirkungsweise", in: Kaminski, Andreas and Gelhard, Andreas (eds.), Zur Philosophie der informellen Technisierung. Darmstadt: Wissenschaftliche Buchgesellschaft, pp. 58–81.

Kaminski, Andreas (2017): "Hat Vertrauen Gründe oder ist Vertrauen ein Grund? Eine (dialektische) Tugendtheorie von Vertrauen und Vertrauenswürdigkeit", in: Kertscher, Jens and Müller, Jan (eds.), Praxis und "zweite Natur" – Begründungsfiguren normativer Wirklichkeit in der Diskussion, Münster: Mentis, pp. 121–139.

Kaminski, Andreas (2020a): Die verwickelte Einfachheit von Vertrauen – und seine spekulative Struktur. Accepted habilitation thesis, Marburg.

Kaminski, A. (2020b): "Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen", in: Wiegerling, Klaus, Nerurkar, Michael and Wadephul, Christian (eds.), Datafizierung und Big Data. Ethische, anthropologische und wissenschaftstheoretische Perspektiven, Wiesbaden: Springer Fachmedien, pp. 151–174.

Kaminski, Andreas (2023): "Die Erfahrung gebrochenen Vertrauens", in: Zeitschrift für Kulturphilosophie 2023(2), pp. 96–115.

Lagerspetz, Olli (1998): Trust. The Tacit Demand, Dordrecht: Springer.

Lahno, Bernd (2002): Der Begriff des Vertrauens, Paderborn: Mentis.

Lauer, Josh (2017): Creditworthy. A History of Consumer Surveillance and Financial Identity in America, New York: Columbia University Press.

Moran, Richard (2005): "Problems of Sincerity", in: Proceedings of the Aristotelian Society 105(1), pp. 325–345.

Moran, Richard (2006): "Getting Told and Being Believed", in: Lackey, Jennifer and Sosa, Ernest (eds.), The Epistemology of Testimony, pp. 272–306.

Moran, Richard (2013): "Testimony, Illocution and the Second Person", in: Aristotelian Society Supplementary 87(1), pp. 115–135.

Nickel, Philip J., Franssen, Maarten and Kroes, Peter (2010): "Can We Make Sense of the Notion of Trustworthy Technology?", in: Knowledge, Technology & Policy 23(3), pp. 429–444.

Ross, Angus (1986): "Why Do We Believe What We Are Told?", in: Ratio 1, pp. 69–88.

Sporer, Siegfried L. (1997): "Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich. Validitätsüberprüfung anhand selbsterlebter und erfundener Geschichten", in: Greuel, Luise, Thomas, Fabian and Stadler, Michael (eds.), Psychologie der Zeugenaussage. Ergebnisse der rechtspsychologischen Forschung, Weinheim: Beltz, pp. 71–85.

Stern, William (ed.) (1903): Beiträge zur Psychologie der Aussage. Mit besonderer Berücksichtigung von Problemen der Rechtspflege, Pädagogik, Psychiatrie und Geschichtsforschung, Leipzig.

Thiebes, Scott, Lins, Sebastian and Sunyaev, Ali (2021): "Trustworthy artificial intelligence", in: Electronic Markets 31(2), pp. 447–464.

Wagner, Gerald (1994): "Vertrauen in Technik", in: Zeitschrift für Soziologie 23(2), pp. 145–157.

Wiengarn, Jörn (2021): Die Grammatik des Vertrauens. Eine Untersuchung in interpersoneller und epistemischer Hinsicht, Köln: transcript.