Philosophy & Technology (2023) 36:66
https://doi.org/10.1007/s13347-023-00664-1

RESEARCH ARTICLE

# Making Trust Safe for AI? Non-agential Trust as a Conceptual Engineering Problem

**Juri Viehoff**[1]

Received: 25 May 2023 / Accepted: 12 September 2023 / Published online: 25 September 2023


**Abstract**
Should we be worried that the concept of trust is increasingly used when we assess non-human agents and artefacts, say robots and AI systems? Whilst some authors have developed explanations of the concept of trust with a view to accounting for trust in AI systems and other non-agents, others have rejected the idea that we should extend trust in this way. The article advances this debate by bringing insights from conceptual engineering to bear on this issue. After setting up a target concept of trust in terms of four functional desiderata (trust-reliance distinction, explanatory strength, tracking affective responses, and accounting for distrust), I analyze how agential vs. non-agential accounts can satisfy these. A final section investigates how 'non-ideal' circumstances—that is, circumstances where the manifest and operative concept use diverge amongst concept users—affect our choice about which rendering of trust is to be preferred. I suggest that some prominent arguments against extending the language of trust to non-agents are not decisive and reflect on an important oversight in the current debate, namely a failure to address how narrower, agent-centred accounts curtail our ability to distrust non-agents.

**Keywords** Trust · conceptual engineering · reliance · distrust · technology · AI

## 1 Introduction

Here are two observations about the theory and practice of trust: First, ordinary discourse on trust has no difficulty with non-human objects and agents. We trust GOOGLE MAPS to provide us with the fastest route to our destination; we say that we don't trust our vacuum robot to properly clean the apartment; and we worry about whether we can trust the battery of our mobile phone while going out for a run. Perhaps unsurprisingly then, a significant majority of human subjects seem unperturbed when tasked with assessing the trustworthiness of artificial intelligence

✉ Juri Viehoff
j.viehoff@uu.nl

[1] Philosophy, Utrecht University, Heidelberglaan 8, Utrecht, The Netherlands

(AI) agents and trusting them (Malle & Ullmann, 2020; Ullman & Malle, 2018). Now contrast this fact with a second observation: The most prominent philosophical accounts of trust depart from paradigmatic interpersonal instances of trust and follow what can be called an 'agent-centered' model. According to this model, paradigmatic or 'real' trust consists in an agent, first, relying on another (with regard to some ϕ[1]), *and*, second, their reliance being grounded in some further interpersonal attitude like good will (Baier, 1986; Jones, 1996), or some moral /normative expectation (Hawley, 2014; Nickel, 2007), or some positive anticipation of responsiveness (Faulkner, 2007; Jones, 2012). Trust, moreover, is seen as the sort of thing that can be *betrayed*, that is, something whose violation renders reactive attitudes fitting (Holton, 1994).

But vacuum cleaners, even robotic ones, cannot have any good will or commitments towards us, nor does it seem reasonable to direct reactive attitudes towards them and feel betrayed: all of these are only warranted when interacting with phenomenally conscious agents, perhaps only when interacting with other humans. So are ordinary speakers open to criticism for their 'loose' speak when it comes to the concept of trust? Against this, a number of recent accounts of trust have defied the tendency to assume that objects of trust must have advanced agential capacities (Owens, 2017; Ferrario, Loi, and Viganò 2020; Nguyen, 2022; Nickel, 2022). The upshot of all these accounts is that, contrary to the most prominent 'orthodox' accounts, the best explanation of trust yields a far more extensive concept according to which non-agential trust, including trust in robots, AI systems, and simple human artefacts, can be just as paradigmatic as interpersonal trust.

This article contributes to this debate. My suggestion is that we can make some progress by conceiving of the question which concept of trust should be operational in interpersonal and human-AI contexts as a conceptual engineering problem: since it is (partly) a matter of choice which extension and intension we give to the concepts we use and how we carve out concepts relative to real world phenomena, we must reflect on what conceptual design choices it would be best to implement. My plan is this: §1 briefly contrasts 'orthodox' agential accounts and recent non-agential ones. §2 describes why it makes sense to think of the question "what counts as trust?" as a conceptual engineering problem. §3 achieves a first step of engineering trust by reconstructing the target concept: To do so, I set up four desiderata (or *design features*) that an account of trust should respect and explain their relevance.

A conclusive analysis of which concept of trust to use is beyond the scope of this paper. Nonetheless, I evaluate, in §4 and §5, a number of core arguments that speak for/against agent-centered or non-agential concepts of trust in terms of how each can help to realize trust's legitimate purposes. Assuming that either account would be fully adopted by a community of reasoners, §4 looks at how each account fares relative to the functional assessment criteria set up in §3. My (perhaps surprising) suggestion is that so long as there is a shared conception and reasoners 'know what they are doing', it does not matter which concept prevails. But, turning to non-ideal conditions (that is, more realistic assumptions that allow for fragmented concept use

[1] In line with much of the literature, I focus here on three-place trust: A trusts B to ϕ.

and discrepancies between what is manifest and what is operative in concept use), §5 argues that prominent arguments against non-agential trust are not decisive. Instead, I note two important oversights in the current debate: First, a failure to address how narrower, agent-centered accounts curtail our ability to *distrust* non-agents. Second, trust's 'therapeutic' or 'proleptic' potential, that is, its ability to shape norms about responsibility and accountability. §6 concludes.

## 2 Setting the Scene: Philosophical Accounts of Trust

### 2.1 The orthodoxy: reliance + agential capacities

Most accounts of trust in the philosophical literature are agent-centered because they conform to what I will call the 'reliance + agential capacities model', or short: *R+AC* (Nguyen, 2022). Let me offer some examples of agent-centeredness amongst the most influential accounts in the literature: Baier, in her pioneering work, simply stipulates that in thinking about episodes of trust we are first and foremost concerned with "one person trusting another with some valued thing" (1986, 236). Likewise, Jones defines trust as a specific *interpersonal* attitude we take towards other people: "At the center of trust is an attitude of optimism about the other person's goodwill" (Jones, 1996, 6). Similarly, Hawley's first pass at trust is that it "involve[es] *two people* and a task" (2014, 2 my emphasis). Approaching trust genealogically, Simpson states that trust in its original variant ('ur-trust') arose out of our need for social cooperation with *other human beings*, for example in child-rearing, or collaboration to secure basic needs like food and shelter (2012, 557). And in their outline of a 'paradigm-based' explanation of trust, Bieber and Viehoff only use examples of trust between humans (Bieber & Viehoff, 2023, 7). What unites these authors is the assumption that it is unproblematic to define their accounts as accounts of *trust simpliciter* rather than accounts of 'trust between persons'.[2]

What I mean by an account of trust being 'agent-centered' then is that it either implicitly assumes, or explicitly defines only objects which have advanced agential capacities to be *trust-apt*. An object is trust-apt if it belongs to the class of objects to which assessments of trust can be correctly (or 'aptly') applied. Nobody doubts, for example, that human beings are typically *trust-apt*: when you consider a person's performance in relation to some task (broadly construed), you can adopt an attitude of trust, or distrust, as to whether or not they will come through. But when the concept of trust is extended to things that lack agential capacities, then, at least according to R+AC, one of two things is going on: A first possibility is that the speaker is using the concept of trust metaphorically, which occurs when the speaker is aware that the 'trusted' object lacks agential capacities. So when, for example, we 'trust' our car's engine to start in the morning, we are really using the concept in a way we

[2] Other authors are more careful: Hieronymi, for example, speaks of 'one version of trust' (Hieronymi, 2008); Faulkner distinguishes between 'predictive' and 'affective' trust (Faulkner, 2007). For a discussion why these are still ultimately 'agent-centered' accounts, see: (Tallant, 2019).

know to be literally false, something we clearly do in other scenarios ("My laptop *loves* to crash while I teach!"). Alternatively, second, the speaker is *anthropomorphizing*[3], that is, genuinely attributing agential capacities to something that in fact lacks them.[4]

Why do philosophers adopt agent-centric accounts of trust? The idea that runs through the most prominent accounts is that trust is best understood as composed of two elements: First, *reliance,* which, as all sides agree, is an action or attitude that we can take towards both agents and non-agents alike ("I rely on the bus to get to work"). Yet, second, when we trust, our reliance coincides with, is grounded in, or motivated by, some further attitude or disposition, one that it only makes sense to adopt in relation to objects that have highly evolved agential capacities. The exact nature of this further elements is, of course, what distinguishes different agent-centered accounts. Very broadly, we can distinguish four kinds of agent-centered accounts according to the further requirement(s) that turn(s) reliance into trust (Nickel, 2017, 196): First, accounts that require some attitude of *optimism* that the trustee will be good-willed (Baier, 1986; Jones, 1996), or have moral integrity (McLeod, 2002). Second, the 'affective/normative expectation' that there is *positive responsiveness* to the trust placed in the agent (Faulkner, 2007, 2011; McGeer, 2008; Pettit, 1995). Third, the idea that reliance needs to be accompanied by adopting a second-personal, normative 'participant stance' towards the trustee (Holton, 1994; Walker, 2006). Fourth, the belief that the trustee has a *commitment* that they will meet (Hawley, 2014), or, relatedly, the idea that the trustee has an *obligation* to which they will conform (Nickel, 2007).

### 2.2 Non-agential accounts of trust

R+AC accounts are by far the most prominent ones amongst philosophers. Recent years have, however, seen a number of new accounts that depart from this orthodoxy. Either explicitly or implicitly, these accounts deny that trust is necessarily agent-centeredness. Some of these accounts have specifically been developed with the aim of providing conceptual resources to evaluate robots and increasingly autonomous AI systems in terms of their trustworthiness (Buechner & Tavani, 2011; Coeckelbergh, 2012; Ferrario et al., 2020, 2021; Grodzinsky et al., 2020; Nickel, 2022; Starke et al., 2022; Taddeo, 2010; Tavani, 2015). Other non-agential accounts stem from a more general dissatisfaction with aspects of agent-centered accounts, e.g.

[3] *Anthropomorphizing* is not, strictly speaking, necessary: the speaker only has to mistakenly attribute those agential capacities needed for trust-aptness (which differ between accounts).

[4] A 'correct' application is one that is neither metaphorical nor based on mistaken assumptions about an object's true capacities. Because the distinction between agential and non-agential accounts tracks trust-aptness (='correct' use), it classifies one seemingly 'non-agential' account of trust as agential: According to Tallant (2019), it is possible to trust non-agents; but it is *mistaken* because it stems from a misattribution of agential features. So for him a 'fully informed' speaker would not attribute trust to non-agents, except metaphorically. Thus, although it 'allows' for non-agential trust, Tallant's account is agential according to my definition.

the absence of a coherently defined 'attitude of trust' (Owens, 2017) or the peculiar explanation R+AC offers of the widespread use of non-agential trust whilst denying other features that are present in all cases (Nguyen, 2022).

As non-agential accounts have been developed with different topics and debates in mind, they differ both with regard to whether they see their proposals as competitors to the orthodox R+AC model[5] and with regard to *how far* they extend trust's reach beyond full human capacities: at least some of those proposing accounts of trust in the context of AI explicitly refer to trust in 'AI *agents*' (Ferrario et al., 2020, 530). Nonetheless, it makes sense to classify some of these under the non-agential label: the understanding of 'agency' that rules in an algorithm because it reliably produces outcomes based on some input variables (one of the examples used by Ferrario et.al.) is very far removed from the more demanding assumptions about agency that figure in all of the various R+AC accounts mentioned. What is noteworthy is that these novel accounts all respect the first side of the equation of what an account of trust entails: they all start with reliance. Where they differ is in terms of the additional feature that distinguishes mere reliance from trust.

Perhaps the non-agential account that comes 'closest' to agential account is the one recently proposed by Nickel in the context of medical AI. His proposal is that we encounter trust when "one entity is disposed to give a second entity discretion over some matter of value on the basis of normative and predictive expectations about that second entity" (Nickel, 2022, 6). Specifically, his idea is that when we combine giving discretion under conditions of vulnerability with an *expectation* that the entity so relied on ought to serve a defined, normatively salient purpose, then we are in the territory of trust. To trust, according to Nickel, is to allow an entity (agent or non-agent) to exercise discretionary authority over something of value to us, and to expect it to discharge its role or function in exercising such discretionary authority. When discretionary authority is transferred to an AI system, say a diagnostic tool in the medical domain, then the system becomes the *object* of trust: trustors *actually*—that is, non-metaphorically—trust the system. However, importantly, whilst the normative expectation (i.e. that the object will serve its function) is directed at the non-agential system or artefact, the moral-cum-emotional element of trust (i.e. attributions of responsibility and blame if the system fails) are directed at its *designer*. Nickel's account is 'closer' to R+AC than those other non-agential accounts to which I now turn because it remains a *normative-cum-moral* account: to trust is, amongst other things, to have normative *and* moral-cum-emotional *expectations*, and, at least the latter, are necessarily agent-directed.

[5] Arguably many accounts that postulate 'new' or 'additional' forms of trust ["e-trust"(Taddeo, 2010); "TRUST*" (Grodzinsky, Miller, and Wolf 2020), ) do not conflict with R+AC: they simply pick out a distinct phenomenon and propose a distinct label. My focus will therefore be on 'general' accounts of trust that are either described as competing or can easily be read as competing with R+AC accounts. Novel forms of trust ("e-trust") give rise to interesting and complex question about the nature and boundaries of concepts that unfortunately go beyond the scope of my analysis (but see remarks on conceptual pluralism in 1.3).

Two other non-agential accounts are distinctly more revisionist: Ferrario et.al. too set out to describe a form of trust that can accommodate AI technology. However, on their conception, there is at least one form of trust ('simple trust') which is "a non-cognitive account of trust based on the concept of reliance" (2020, 530) according to which X trust Y when "X is willing to rely on Y to perform an action A pursuing a goal G, and X plans to rely on Y without intentionally generating and/or processing further information about Y's capabilities to achieve G" (2020, 530). What is notably about this account is that the additional requirement beyond reliance they formulate, namely that the trusting agent plans without controlling or monitoring Y's performance, is so weak that the concept of trust can easily be extended beyond agents: though their primary goal is to provide an account that allows for the assessment of AI in terms of trust(worthiness), we may adopt this attitude in a much broader set of cases including much simpler artefacts.

A third account of non-agential trust by Nguyen is the most revisionist in that it explicitly extends trust-aptness to cover simple artefacts like ropes and even non-artefacts like the ground beneath our feet (2022, 217). According to this account, paradigmatic trust always consists in reliance paired with what he calls an "unquestioning attitude" on the side of the trustor. He describes this attitude as follows: "to trust something (..) is to put it outside the space of evaluation and deliberation — to rely on it without pausing to think about whether it will actually come through for you." (2022, 214). Specifically, we have adopted this attitude when we have a first-order disposition to accept 'immediately' that a trusted entity will ϕ *and* a second-order disposition to 'deflect questioning' our first-order disposition (2022, 225). Once we have this complex attitude in full view, claims about trust in non-agents being either metaphorical or mistaken appear much less plausible: we can take this unquestioning attitude to non-agents and, when it is paired with reliance, we instinctively deploy the concept of trust for such cases.

### 2.3 A genuine disagreement?

Let me close this section by addressing a worry: In spite of how I have presented agential and non-agential accounts, namely as *competing* explanations of the nature of trust and hence of what our concept should be, one could think there really is no disagreement here. After all, could defenders of agential and non-agential accounts of trust not each happily agree that the other has identified an important form of trust? A first indication that spells doubt on this seemingly conciliatory response is that defenders of R+AC accounts have started to argue *against* extending the language of trust to non-agents. At least those authors who reject non-agential trust and warn against their use must see a genuine dispute about what our 'real' concept should fundamentally track (Al, 2022; Budnik, 2018; Hatherley, 2020; Ryan, 2020).

Of course, these authors may be mistaken. But there is also a second, more principled point why pluralism is hard to countenance. Advocates of agential and non-agential accounts both aim to establish a robust difference between trust and reliance—and the central features of trust they propose are meant to track this

distinction. For agential accounts, accepting that there is another form of trust for which the trust/reliance distinction operates differently (say Nguyen's unquestioning attitude account) raises the danger of rendering their way of drawing the distinction between trust and reliance in *interpersonal* cases ambiguous, namely when the trustor has adopted the unquestioning attitude, yet lacks the further agential-capacity-related conditions required by R+AC. Conversely, accounts that draw the trust/reliance distinction in terms of non-agent-centered features run into trouble when agential features of trust are present but non-agential ones are not.

To illustrate, consider one of the most famous examples in the literature to account for the difference between trust and reliance: the putative tendency of the people of Königsberg to rely in some aspect of their daily life on Kant's clockwork-like regiment of taking a walk at a particular time of day. Since Baier (1986, 235), this example has been used to establish that one can rely interpersonally *without* trusting. But according to those defending non-agential accounts, the people of Königsberg *would* trust Kant if their reliance became unthinking or they stopped monitoring, whether or not they attribute good will, commitments etc. In relation to non-agential accounts, Gordon (2022, 568) has recently described cases where an agent seemingly trusts *whilst* monitoring compliance, which amounts to trusting *without* having an unquestioning attitude. If these are indeed instances of trust, then this would show that an inverse problem arises for some versions of non-agential trust: there would be interpersonal cases where the agent trusts on the agential account but 'merely relies' on at least one version of the non-agential picture.

Given these contradictory assessments of identical cases, it seems hard to deny, at least on a 'classical' understanding of the structure of concepts and their explanation in terms of necessary and sufficient conditions (Margolis & Laurence, 2023), that agential and non-agential accounts propose competing concepts of trust: they differ in their respective extensions and intensions, and each carve out trust differently relative to other concepts, most notably in relation to 'mere' reliance.[6]

[6] An anonymous referee helpfully points out that trust may be internally complex, consisting of a prototype/paradigm or plural structure: whether some instance counts as an episode of trust depends on whether it shares some elements with a 'prototype', and hence trust is a matter of family resemblance, but there are no necessary features. Or, trust may be a pluralist or multivocal concept, that is, composed of different prototypes or paradigm instances, each with their own central elements (Weiskopf, 2009). Trust *between persons* could then always have to go beyond an attitude of merely relying unquestioningly (by adopting one of the attitudes specified by R+AC accounts), yet trust in non-human agents and artefacts (though equally paradigmatic), could occur when the trustor stops monitoring, whether or not they also display the more complex attitude required in episodes of interpersonal trust.

If trust displays such a pluralist structure then this may pose two challenges: First, for a pluralist concept, it may be difficult to engage in conceptual engineering in any fruitful way for lack of a 'target concept': it might simply be impossible to specify what the concept should help us to do. A second problem could be that, even if we can meaningfully 'engineer' trust, we have to broaden the set of possible candidates to consider options that go beyond univocal explanations. Thus, we would have to include multivocal, pluralist renderings of trust amongst candidate explanations. I agree with the referee that any ultimate assessment would need to consider the option of engineering a pluralist concept of trust. But for reasons of space and to reduce complexity, I here focus on the proposals that have been put forward by different authors.

## 3 Why conceptual engineering?

Two claims constitute the point of departure for conceptual engineering: First, that our operative representational devices, including our concepts, do not only reflect our practices but can make some theoretical or practical difference in the world. Second, the claim that it is, at least up to a point, within our individual and/or collective control what representational devices, including concepts, are operative amongst us. As a result, we should not ask—or at least not *exclusively* ask "What concept of trust is implicit in our linguistic or historical practices?". Instead, we should (also) ask: "What concept of trust would it be best to have in operation?". With this question in focus, conceptual engineering is the attempt to systematically develop criteria for making conceptual choices by mapping out our options and reasoning about what concept it would be best to have.[7]

Approaching trust as a conceptual engineering problem changes the parameters of the debate: Whether our aim is to defend orthodox agent-centered or non-agential accounts, our *justificatory strategy* for doing so now needs to be based on considerations that are quite different from arguments that appeal to the intuitive force of particular trust-related judgments. For example: in ruling out the applicability of the concept of trust to AI technology, say machine learning in medical diagnostics, we could no longer simply appeal to the intuitive judgment that we are here not dealing with trust but only 'mere' reliance, or putatively evident conceptual truth about the non-applicability of trust(-aptness) to non-human entities.[8] Instead, we will have to show *why* parceling our conceptual landscape in such a way as to rule out trust's applicability to AI technology would be more advantageous than a competing conceptualization according to which assessments of trust can non-agentially be applied to these technologies.

Before I get deeper into the 'how' of conceptual engineering, I want to note two reasons why conceptual engineering seems especially worthwhile for the concept of trust: A first reason simply flows the conceptual landscape of trust and related concepts. Put simply: we have options. It is *prima facie* possible to alter our conceptual practice in line with the theoretical explanation of what the concept of trust should be. There are several concepts in trust's vicinity that we could use as auxiliaries if *sharpening* trust (e.g. restricting its extension to full agents) is what we should do. For example: we may refrain, upon reflection, from talking of trust in non-agents like tools and artifacts, instead exclusively utilizing the concept of reliance in this

[7] Conceptual engineers must of course also defend these initial claims. I make no attempt to defend conceptual engineering as a method *in general*. Rather, and in spite of important worries and powerful criticisms raised in philosophy of language and metaphilosophy, I will simply accept it here as a plausible approach in order to see what kind of consequences follow if we apply it to a particular domain of theorizing. (For contributions to this rapidly growing field, see (Burgess et al., 2020; Capellen & Plunkett, 2020; Cappelen, 2018; Koch, 2018; Nado, 2021; Thomasson, 2021).

[8] "Although well intentioned, applying trust to AI is a category error, mistakenly assuming that AI belongs to a category of things that can be trusted." (DeCamp & Tilburt, 2019, 390)

area. And we may recommend this usage to others with the aim of shifting our conceptual practice in this direction.[9]

A second reason turns on our current predicament. It derives from the conjecture that conceptual engineering is particularly feasible and important when rapid social, political and technological change causes shifts in a concept's context of application, paired with the factual claim that we are presently witnessing such a shift.[10] Is there any evidence that supports the factual claim that we are in a 'conceptual transition period' with relation to trust? One reason to think so is that due to the magnitude and speed at which technological change occurs, 'trust claims' are now frequently made in relation to novel types of agents and artefacts (ranging from autonomously driving cars to search algorithms) whose characteristics are difficult to assess. For example, one recent experimental study indicates that the use of digital technologies and AI seems to cause shifts in how users understand their concepts: young adults—more accustomed to a world in which many decisions around them are made by algorithmic systems—have been shown to be more likely to attribute agency and responsibility to autonomous- (seeming) technology than older subjects (Kneer, 2021; Stuart & Kneer, 2021).

Why is conceptual engineering especially feasible and important under these conditions?[11] Take feasibility: conceptual innovators may stand a better chance of successfully implementing their novel concepts when there is a receptive audience for conceptual innovation. Whether the audience is receptive in turn plausibly depends on how well the present concept serves those who use it. Yet stemming from misunderstandings or otherwise suboptimal outcomes, a concept's 'use value' under conditions of rapid conceptual change through technology may be of limited value for communication.

Turning to the significance of conceptual engineering projects, it seems that systematic philosophical reflection on the concepts we use and their alternatives becomes especially urgent when there is ideological pressure towards the acceptance of particular renderings of our concept(s), something we might encounter when prevailing concept use is shifting. The discourse on AI and trust is a good example of this: several theorists have worried that extending the concept of trust and trustworthiness to cover AI, e.g. to 'social robots', search algorithms and AI-chatbots, is in fact a sophisticated ideological project to foster positive attitudes towards such products and their implementation; reduce healthy skepticism about their safety and usefulness, and, most nefariously, to 'naturalize' these attitudes of unquestioning

[9] Of course, conceptual engineers cannot simply alter the intension/extension people have internalized in their concept use. But we can, to some extent, shape the use by making proposals for changes and 'normalizing' the preferred ameliorations. I unfortunately lack the space to discuss in detail how the implementation of an engineered concept for trust would work out. (But see §6 for some reflection about more or less successful implementations). For a plausible 'optimistic' view on how conceptual engineering can be successful, see: (Simion & Kelp, 2020).

[10] (Himmelreich & Köhler, 2022) make a similar claim about AI and the concept of responsibility.

[11] On the relationship between evolving technology and conceptual engineering, see: (Veluwenkamp et al., 2022; Veluwenkamp & van den Hoven, 2023)

acceptance by embedding them in our linguistic and conceptual practices [(Al, 2022; Hatherley, 2020), see §5]. Making 'trust'-related discourse more salient in relation to such technology also de-emphasizes practices that make trusting them unnecessary, namely *monitoring* and strict regulatory *oversight* (Bryson, 2018).

## 4 Trust: Designing the Conceptual Engineering Task

We can distinguish the tasks of conceptual engineering into four broad phases: 'description', 'assessment', 'improvement', and 'implementation'. First, we must understand what *exactly* it is that we are in the business of engineering: having some description of the subject matter is an essential intial step for determining the target of a design project (Isaac et al., 2022, 3). Moving on to the second, 'assessment' component, we need to get a clearer idea of what we want from our concept. Thus, we (a) set out what we want the concept to do (or not do) and we then (b) assess existing conceptions of the concept in terms of whether they meet these criteria. But what considerations determine what goes into (a)? It is helpful to distinguish between a project's general *purpose* and its specific *goals* (Isaac et al., 2022, 4–5). Purposes are our end-state objectives, whereas *goals* are the more narrowly confined means by which we intend to realize our purposes. So, for example, the purpose of the ameliorating project for 'woman' (Haslanger, 2012, 2020b)—is to advance social justice and to end gender-based oppression, whereas the goal is to change the truth-conditional content of the concept of 'woman'.

If our aim is to improve the concept of trust, we inevitably have to start with some understanding of that concept (i.e. the initial, descriptive stage of the process). But how may we do this without already adjudicating between competing agential and non-agential explanations? Perhaps the most natural thing would be to look at some central, undisputed cases. But the problem with this approach in relation to trust is that the conceptual dispute we described earlier revolves around whether *undisputed* central cases (interpersonal trust) are the *only* central cases (and of course, *why* they are).

Another way of identifying the target concept may initially seem more promising, namely one whereby we start by identifying, in broad terms, the *function* of the concept we are investigating only to then assess which conceptual rendering best helps us to realize this function. There is perhaps evidence that some philosophers writing on trust have embarked on explanatory projects not too dissimilar[12]: Jones, for example, has set out to account for trust by first offering a 'job description' for our concept of trust and trustworthiness (Jones, 2012, 62). And Simpson (2012) embraces a related strategy, namely one that is *genealogical*: What we should do

[12] I say 'perhaps' here because both Jones' and Simpson's approaches are, it seems to me, best understood as explanations of the function of our practice of trust, whereas the conceptual engineering project is distinctively interested in describing the function of our *concept of trust*.

in trying to come up with a definition of trust is to try to understand what purposes those practices demarcated by the term 'trust' have had in human history.[13]

The problem here is that we face a similar difficulty of offering a description (this time of trust's central *function*) that is neutral between agential and the non-agential accounts: The genealogical story Simpson and Jones tell is that, as social beings that fundamentally depend on one another, we needed practices that allowed us to effectively and efficiently manage our dependency *on other people*. (And, to relate this to conceptual engineering, we needed concepts that demarcated specific ways of voluntarily depending on other people, and concepts that demarcated responding to other people's recognized voluntary dependence on us.) Trust's function on both their accounts is to allow us to manage this form of *interpersonal* dependency, and so whether x qualifies as an instance of trust depends on whether or not x discharges this interpersonal function. Against this, Nguyen has simply claimed that we should understand trust in terms of a broader function: "The basic form of ur-trust, I'm suggesting, is *agential integration*. Trust (…) involves the attempt to bring other people and things into one's agency, or of joining with other people and things into collective agencies." (2022, 241). Each of these functional stories lead to plausible initial descriptions of trust.

Is there a way to get around this difficulty? My suggestion is motivated by an observation regarding the way in which accounts of trust are justified in the philosophical literature. Although much existing philosophical treatment proceeds by reflective equilibrium between individual case judgments and rules for what counts as trust, there is also an implicit additional explanatory strategy that runs through the literature, namely one according to which convincing accounts of trust must satisfy a number of *structural* constraints or desiderata. Though not all of them are always used or made explicit, four criteria strike me as very widely shared amongst philosopher writing on trust:

(1) **Reliance**. Perhaps most importantly, any convincing account of trust must be able to explain how trust *differs* from nearby elements in the conceptual landscape. Most prominently, any account of trust worth its salt must be able to distinguish between genuine *trust* and ('mere') *reliance*.[14]
(2) **Rationality/Explanatory Power**. Any account of trust should meet an explanatory desideratum in that it must contain—or at least be compatible with—a plausible story about how trust can be rational for an individual and, derivatively, how trust can explain social cooperation (Nickel, 2017).
(3) **Affective/Emotional Responses**. An account of trust should be able to account for the distinctive affective/emotional response characteristic of central episodes: when trust fails, we feel not just disappointed, but *betrayed*.

[13] For a critical discussion of the differences between 'functionalist' or 'paradigm-based' explanations and genealogical approaches, see: (Fricker, 2019).

[14] Cf. (Hawley, 2017, 233): "Vindicating this distinction [between trust and reliance] has been regarded as an essential criterion of success for accounts of trust (…)."

(4) **Distrust**. Accounts of trust should explain trust's relation to distrust: if our account cannot make sense of the threefold conceptual structure that splits our practice into trust/distrust/non-trust, it would be defective in that it only captures one part of a practice that needs to be understood holistically (D'Cruz, 2020; Hawley, 2014).

Crucially, these desiderata (or at least most of them) are not only accepted by those defending the R+AC orthodoxy: Nickel, Ferrario et. Al. and Nguyen, for example, seek to establish that their non-agential trust too can distinguish trust from reliance; and Nguyen at least also addresses the issues of betrayal and his account's ability to explain distrust (Nguyen, 2022, 228). My idea then is that we take these four desiderata to be non-prejudicial, preliminary assessment criteria for what it is that our concept of trust should aim to do.

To be clear, we have, by setting out these criteria, to some extent merged the 'descriptive' and the 'assessment' stage of the engineering framework. I do find this a promising strategy: on the one hand, it ensures that proposals for conceptual amelioration display a continuity to what is generally taken to be trust's descriptive conceptual core. This is helpful because, if an amelioration proposal meets (most of) these desiderata, it cannot be dismissed based on the charge that those proposing it are changing the topic (Cappelen, 2018, 98; Haslanger, 2012, 225). Yet on the other hand, judging explications of trust based on these preliminary desiderata leaves much room for conceptual improvement.

There are two reasons for this: First, conceptual innovators can appeal to a variety of ways of understanding these structural constraints. For example, all non-agential accounts offer explanations of the difference between trust and reliance that are grounded in very different considerations that those typically made by R+AC accounts. One is therefore not beholden to a particular understanding of the nature of trust by respecting them. Second, my suggestion is that, as conceptual engineers, we consider these constraints to be *preliminary*: if it turns out that rejecting any of them would lead to a concept of trust that either satisfies the other desiderata much better, or a conception that realizes some additional important practical or theoretical purpose, then this would support the case for rejecting this desideratum.

The crucial assumption in the background is that these desiderata must *themselves* be justifiable from a conceptual engineering standpoint. An analogy to actual product design may be useful here: Suppose we aim to build a technological artefact, say a novel kind of wireless earphones. To do so, we initially stipulate a set of specifications, say in terms of size, connectivity, sound volume, and sound quality. But moving towards constructing a prototype, we perhaps come to understand that some specifications are not jointly realizable. Moreover, we may find that some additional features we had not considered—say in-ear fit and weight—are equally or more important than our original specifications. We would then revise the original set of criteria. Now some criteria—e.g. connectivity—are highly unlikely of dropping from the list if what we are in the business of designing is a pair of earphones. But even essential characteristics like these are fundamentally determined and justified in functional terms. So my proposal is to assess (a) how well different accounts

each can meet the four desiderata just outlined and (b) whether they would advance additional practical and epistemic purposes we identify along the way.

Since my focus in this article is on one crucial 'feature difference' that separates prominent accounts of trust (agential vs. non-agential), my suggestion is that, when we turn to the issue of assessing our criteria against existing explanations, we can dramatically simplify the exercise by shifting from what we may call an 'open competition' to a two-way comparison between agential and non-agential conceptions.

## 5 Assessing R+AC vs. Non-Agential Accounts: Ideal Theory

In relation to the topic at hand, the conceptual engineering standpoint directs us to the following question: "Would it overall be a positive or a negative thing to parcel the conceptual space in such a way that the extension of the concept of trust covers both agents and non-agents?" Having developed a set of assessment criteria, we can now more specifically evaluate how agential/non-agential accounts fare in terms of meeting the desiderata of (i) offering a useful trust-reliance distinction, (ii) furnishing an account that renders trust explanatorily useful, (iii) explaining reactive emotions associated with violations of trust, and (iv) accounting for distrust.

Before we engage in this exercise, let me propose one distinction that will structure this and the following section. In this section, I want to compare agential and non-agential accounts under conditions of what I will call *ideal-theory*, before moving on to non-ideal considerations in the next one. I mean this distinction of ideal/non-ideal to track something roughly analogous to what Rawls (2001, 13) suggests in relation to conceptions of justice, except that we are of course dealing with partial vs. full-compliance amongst *concept users*. In this section I reflect on which concept would be more suitable to have *if* all concept users followed one or the other account in full and were aware of this, before considering complications that arise if we are reflecting on what to do when some reasoners will fail to (fully) adopt either account.

### 5.1 The trust-reliance distinction

Being able to properly separate trust from mere reliance is widely seen as a criterion that any successful account of trust must meet (Goldberg, 2020). Nonetheless, as conceptual engineers, we should ask *why* it should matter for an explanation of trust to be able to offer a plausible distinction between these concepts: what functional purpose is served by drawing such a distinction? The most plausible answer is that having the distinction serves us to orient ourselves practically when we need to manage vulnerability. Take a simple sentence like "I rely on x, but I wouldn't trust x!" Whether x is the slightly tipsy friend that is our last ride home or whether x is the shaky wooden bridge we must cross to escape the hungry tiger, formulating sentences/thoughts in this way seems important for us as practical agents, for it allows us to distinguish an attitude-action of merely planning to make use of an agent or object from something more involved.

R+AC accounts, we saw, believe that this 'something more' should derive from our distinctly human responses and mechanisms of managing dependency on other people. Take the 'positive affective responsiveness' variant of R+AC: here, the concept of trust tracks affective responsiveness because doing so helps us to focus on this element in our practical deliberations as agents. Al, drawing on Jones, suggests that "trust is not merely the acceptance of these dependencies; it also helps us to (partly) overcome these dependencies" (2022, 8). What he has in mind is the fact that other human agents, but not artefacts, can be motivated through our placing trust in them. So if the practical purpose of our concept of trust (in counter-distinction to reliance) is to understand how we can motivate our agential environment to get what we want, then this way of drawing the distinction seems quite convincing. And it makes trust almost necessarily agent-centered.

So does this plausible rendering of the reliance/trust distinction demonstrate that non-agential accounts are inferior from a functional perspective? One might think so for the following reason: at least according to some of the non-agential accounts described earlier, we lose the ability to draw a practical distinction between trust and reliance in paradigmatic interpersonal cases: Recall how I suggested that non-agential accounts end up classifying the people of Königsberg as trusting Kant when they rely on him without questioning their doing so (Nguyen) or monitoring his performance (Ferrario et. al.). *Prima facie*, this seems to vindicate R+AC accounts.

But upon closer inspection, this turns out to be illusionary. After all, we would only have a decisive reason for R+AC on this criterion if defenders of non-agential accounts could not offer a rendering of the trust/reliance distinction that would be equally practically significant for us as agents. And this, it seems to me, is precisely what they are in fact able to do. Nguyen's discussion of the trust/reliance distinction in terms of an unquestioning attitude is the most sophisticated defense here: Far from being unable to offer an account of the trust/reliance distinction, the unquestioning attitude account provides us with cues that too are of central importance to our practical agency. When you rely on someone's or something's x-ing, you "act on the supposition that she will x" (Hawley, 2014, 4; Holton, 1994; Goldberg, 2020). As Nguyen explains, understood this way relying on a person or object is quite compatible with constantly checking whether they will come through. Yet within our practice of relying, there clearly are cases where we stop monitoring and rely with the unquestioning attitude. As explained earlier, Nguyen thinks of this special form of reliance, unquestioning reliance, as one essential way of coping effectively as agents: we would simply be less-than-optimally functional if, in planning our life, we would never take an unquestioning attitude towards aspects of our environment. Put differently, relying whilst taking this attitude (i.e. trusting) allows to enhance our practical agency (2022, 231). Yet of course, relying unquestioningly also gives rise to particular forms of vulnerability and directs us towards important practical questions: "On what kind of entity should we unquestioningly rely?"; "Which forms of agential outsourcing render us too vulnerable?" The upshot is that at least this way of drawing the trust/reliance distinction also track something of fundamental importance to us as agents.

It seems clear then that we cannot find some *definitive* practical advantage in either the 'affective responsiveness' or the 'unquestioning attitude' rendering of the

distinction that I have analyzed. Though lack of space prevents me from contrasting different agent-centered and non-agential accounts, my suspicion is that these two being arguably the most elaborate explanations of the trust/reliance distinction from a functional perspective, there is no decisive reason to prefer R+AC over non-agential accounts based on considerations of orienting us in the world.

### 5.2 Superior Explanatory Potential?

When the International Astronomical Union (IAU) in 2006 adopted a definition of PLANET according to which Pluto no longer counted as a falling under this concept, the decision was reached based on the explanatory power of this rendering of the concept.[15] Trust too is often understood to be explanatory at the individual and the collective level: "*because* P trusted Q" can be a good explanation why P acted; similarly, "*because* they trust each other" can be an explanation of social cooperation. Can our aim of choosing a conception of trust that is best placed to live up to this explanatory role help us to adjudicate between agent-centered and non-agential accounts of the concept? My suggestion below is that though an important consideration, it is very difficult to come up with some clear overall judgment.

Reflecting on how explanatory potential might help us to adjudicate between narrower and broader conceptions of *interpersonal* trust (where 'narrow' and 'broad' are cashed out in terms of the set of motivations based on which an agent can act and count as 'trusting'), Nickel concludes that broader conceptions seem to have an edge in relation to two conditions he calls the *Explanatory Constraint* on trust. According to this constraint, a conception of trust should "(a) be explained as the outcome of central concerns or interests of the relevant actors, and (b) explain the emergence and sustenance of cooperative practices and social institutions" (2017, 197). Accounts of trust that define trust in terms of demanding motivational criteria (for example an expectation of good-will by the trustee, or that the trustee is trustworthy in a rich sense) fare *worse* than more permissive accounts—frequently used in empirical disciplines—according to which exclusively strategically-motivated behaviour can also count as trust. These latter accounts, for example (Hardin, 1996), better meet both (a) and (b) because "under a single concept, so to speak, the unrestrictive view allows for the emergence of stable patterns of cooperation from both strategic and non-strategic dispositions towards reliance." (Nickel, 2017, 199).

Now the question we are facing here is different in that we are trying to understand whether a narrower or broader definition of *trust-apt* entities provides more explanatory potential. And it seems to me that there is no parallel strategy for establishing better explanatory fit for this case: The problem is that each side can appeal to a number of factors according to which their understanding of trust-aptness is more explanatorily powerful. Unfortunately, I lack space to fully develop all of the

[15] The reason Pluto was demoted to 'dwarf planet' was the requirement that a planet, besides orbiting a sun and having a round shape, must have "cleared the neighbourhood around its orbit" (International Astronomical Union, 2006). The requirement was justified on considerations stemming from the best theories of planet formation. (Soter, 2006).

relevant considerations here, so a sketch of the kind of argumentative exchange will have to do:

Take the example of agent-centered accounts that conceive of trust as reliance plus 'affective responsiveness' or normative expectations: if we understand trust in terms of relying *whilst* anticipating that the trustee will affectively respond to our trust in them, then we have a very compelling explanation of how trust generates social cooperation: it 'uses' our affective responses to effectively prompt others (who can understand them) to become motivated to do what we expect them to (McGeer & Pettit, 2017; Pettit, 1995). Likewise, normative expectation accounts can point to how the desire to live up to normative expectations or to keep one's commitments motivates people to do what is necessary for cooperation. By contrast, defenders of such accounts might suggest, the fact that people rely on others whilst adopting an unquestioning attitude does very little to explain why trust is rational and leads to social cooperation.

The response by defenders of the unquestioning attitude account will likely be threefold. First, they can claim that even if the agent-centered story is *more* compelling, this does not *per se* say anything against the power of non-agential accounts: If both agent-centered and non-agential accounts would in fact yield equally rich explanations of how social cooperation arises, this would be a strong parsimony-based reason *against* agent-centered accounts: after all, R+AC accounts start with a much richer description of what trust-apt entities must be like. Second, they will claim that non-agential accounts too go some way towards explaining how social cooperation arises: when, in the course of relying on others, we stop monitoring whether trustees act as we require them to, then we thereby 'settle' on a specific course of action. Having robust dispositions of this kind is just as important in explaining how social coordination becomes possible within groups as the use of our affective/normative system to bolster positive responsiveness to our trust.

The second, more combative response will be to question why *social cooperation* in particular should be the primary target of trust's explanatory potential: why not, instead, or at least alongside this target, consider the explanatory potential that our account has in relation to a wider class of human practices, namely those through which we intend to improve our abilities to pursue ends by means of extending our practical agency (Nguyen, 2022)? When it comes to explaining this wider class of practices, the disposition to stop monitoring provides a powerful explanation: whether in unquestioningly relying on other people or our mobile phone, our ability to trust in this sense explains how we enhance our agency through 'outsourcing'.

### 5.3 Accounting for our affective/emotional responses

It is commonplace in the philosophical literature on trust that reliance and trust give rise to different emotional responses when the agent on which we relied or trusted does not come through: when we relied, we may feel at most disappointed; when we trusted, we can also feel betrayed (Holton, 1994, 66; Hieronymi, 2008, 215; Goldberg, 2020, 98; Brennan, 2021, 775). Whilst trying to grasp trust on a traditional conceptual analysis picture takes this as a simple datapoint that an account of trust

must respect, the conceptual engineering perspective again directs us to ask why the best version of the concept of trust should respect this datum.

I think the best thing to say here is that, whether we like it or not, the nature of the practices to which the concept of trust must be responsive has some implicit inferences attached to it so that any attempt to propose an account of trust according to which trust is disconnected from something like the disposition to feel betrayed when it is violated would amount to proposing a different concept. Put differently, we would be changing the topic if the concept we propose ignored completely those emotional/affective attitudes that typical come with violations of trust.

But does this not settle the dispute between agent-centered and non-agential accounts in favor of the former? After all, the language of betrayal really only seems non-metaphorically applicable to agents with evolved agential capacities: we can only (aptly) feel betrayed when another agent has in fact betrayed us, and, so it seems intuitively, in order to betray somebody one needs phenomenal consciousness, intentions, and, importantly, motives. However, the issue is significantly more complex than it first appears. Here is how non-agential theorists of trust might mount a defense: A first response is that we can feel (something like) betrayal when our trust in a non-agential artefact or system is violated, at least in those cases where the artefact or system has been designed by a responsible agent with a purpose, and the trustor has reason to feel betrayal because of the normative failing of the human beings that designed the artefact (Nickel, 2022). I do not think, however, that this response is successful: after all, the R+AC response will likely be that when the betrayal is directed at the designer of an artefact, then this too is (or should be) where our trust is aimed at. When you feel betrayed by the engineer that designed your phone, then what you were trusting were the engineer's abilities, competence, good-will etc., but *not* the phone.

A second, more revisionist response starts with a demand for explanation of the relevant notion of 'betrayal' appealed to by defenders of R+AC accounts: Notice, first, that not all unsuccessful episodes of interpersonal trust merit the whole set of emotions associated with 'betrayal'. For example, when a student that I trusted to attend does not show up for a zoom office hour, I feel something other than disappointment— I feel slightly annoyed. But 'betrayal' with its rich emotional undertones seems far too strong a word here. So even for R+AC accounts there must be a special 'technical' sense of betrayal at work rather than the standard, full-blown emotional one. This raises a problem for R+AC accounts once we look at our emotional responses when *objects* that are deeply integrated in an unquestioning agent's practical agency (say a violinist's bow during a concert) fail: here too, the agent's emotional reaction goes beyond mere 'disappointment' (though it probably stops short of betrayal). There is a real sense of 'being let down' by something close to us, and we experience a sense of horror at the sudden violation of our expectations and the unexpected disintegration of our extended agency.

So if trust is not strictly speaking related to our full-blown understanding of the emotion of *betrayal*, then what is it connected to? Though the issue merits more detailed analysis then I can afford it here, we can make at least some progress by trying to distinguish 'trust betrayal' as an emotion from other phenomena in its vicinity. Take the phenomenon of resentment towards somebody who failed to meet a

commitment towards us. Clearly, not all instances of resentment are instances of betrayal: when you already expected somebody to be a terrible person, you will unlikely feel betrayed when they act as you expected them to. What jumps at us is that betrayal, unlike resentment, necessarily contains an "element of surprise" (O'Neil, 2012, 308). Similarly, what unifies 'full-blown' betrayal and those less heated attitudes that come out in more minor violations of trust really is an unexpected reversal of the attitude we had adopted.

The final step in the argument is to suggest that this element of surprise, rather than the full-blown set of 'hot' reactive attitudes of feeling betrayed is what is an immovable aspect of our concept of trust. But, as we saw above, we do feel 'betrayed' in this lesser, more technical sense of experiencing a sudden reversal of reasonable expectations in relation to non-agents just as much as we do in relation to agents (and, we do not always feel the full force of betrayal, even when our trust has been violated in interpersonal settings).

### 5.4 Accounting for Distrust?

One of Katherine Hawley's significant contributions to the philosophical reflection on the subject lies in making explicit the broader conceptual landscape in which 'trust' is embedded: We are bound to fail to correctly characterize important features of trust, Hawley thinks, if we fail to pay attention to what trust stands in opposition to, namely *distrust* (rather than the mere absence of trust). Moreover, failing to reflect on distrust may cause our account to lack an appropriate explanation of why trust is frequently not the appropriate attitude to take towards the behaviour of others. Hawley's own account, which I will take to be the most developed argument for how reflecting on distrust favours agential explanations ties trust-aptness to an agent's commitments: A trusts B just in case A holds that B will live up to a commitment B has. If we extent this account to distrust, A distrusts B when A holds that B has a commitment to ϕ but chooses not to rely on B living up to this commitment.[16] Beyond offering an elegant explanation of the nature of trust/distrust, the account also offers a plausible story about cases where neither trust nor distrust are appropriate, namely those where B simply has no commitment to ϕ (Hawley, 2014, 4).

Is there an equally plausible explanation of the trust/distrust relation available to non-agential accounts? One strategy might be to extent the language of commitment and obligations to non-human agents: Fricker (2023), 8) suggested that Hawley's inclination to treat institutional (dis)trustworthiness and institutional (non)-reliability (Hawley, 2017, 4) synonymously should be resisted. Institutions and other non-human agents can have commitments and obligations, and when they do, the register of trust is in order. But even if successful, this would only indicate that some

[16] Can A distrust B to ϕ and still rely on B's ϕ-ing? If, as I suggested, to rely on B's ϕ-ing is to plan and act on the supposition that B will ϕ, then probably not. ("Probably": What A can of course do, when distrusting B to ϕ, is to rely on B not ϕ-ing. If relying on B not to ϕ is a form of reliance on B's ϕ-ing, then one can rely whilst distrusting. But it probably isn't).

non-human agents—those that can have commitments, fit the bill. This would still leave much of the actual (dis)trust-related common discourse unaccounted for and, more importantly, it would not really connect non-agential accounts to the issue of distrust in a systematic manner.

A more robust defense is this: the 'more' that turns reliance into trust on the non-agential picture, recall, is the 'putting the issue of whether or not the trusted object or agent will perform '*to the back of one's mind*'. If this is the case, then distrust can be partly described as the negative opposite of this phenomenon. This can occur in one of two ways: First, one can refrain from relying with the unquestioning attitude by simply deciding not to rely: our distrust in x often manifests itself by choosing not to rely on x. But not all episodes of deciding not to rely are instances of distrust (Hawley, 2014, 4). Similarly, one can plausibly distrust x even when relying on x. This occurs when one relies on x and yet one adopts a stance towards performance characterized by a robust disposition to constantly track whether or not the agent or object will come through (Ferrario et al., 2020, 531). So what characterizes distrust on the non-agential picture is the attitude we take towards the agent or object of (dis)trust: one *questions* whether x will or would perform as expected or intended (Nguyen, 2022, 229).

How exactly this distrusting attitude manifests itself will depend on the specific situation: when one need not rely on x, distrust may simply register as a disposition not to rely paired with a belief that one should not rely (unquestioningly). But when one distrustingly relies on an agent or object, perhaps for lack of a better alternative, one is likely to experience the urge to constantly monitor described by Ferrario, Loi, and Vignanò. And where one cannot monitor performance, perhaps due to vulnerability or power asymmetries, one frequently experiences anxiety and related emotions.[17]

If this is our understanding of distrust, then non-agential accounts do offer practical guidance: reflecting on whether to distrust helps us to decide whether we ought, rationally, to 'lower our guard' or continue to monitor vigilantly. And, like the agential understanding of trust/distrust in terms of commitments, this explanation of distrust also offers a plausible story about what we are doing when we are neither trusting nor distrusting: "to non-trust is to have a neutral attitude, which is entirely open and unresistant to questioning and non-questioning, as the situation suggests." (Nguyen, 2022, 229). Analogously to what we discovered before then, it does not seem that either agential or non-agential accounts track something fundamentally more important from a functional perspective when it comes to accounting for distrust.

### 5.5 Summary

My conclusion for this section may be surprising: It is by no means obvious that either agential or non-agential accounts fare better in relation to the desiderata or design specifications we set up in §3. I think there is a deeper lesson in the offing

[17] I thank an anonymous referee for prompting me to clarify and better describe the non-agential explanation of distrust.

here: Agential and non-agential accounts each come as set of 'bundled' claims that tend to mutually support each other. What I mean by this is that explanations of how each account meets the four desiderata are linked so that, for example, the trust/reliance distinction proposed by most agential accounts is what it is in part *because* of a particular rendering of how we must understand the notion of betrayal when the trustee does not come through. Non-agential accounts deny that trust must always trigger anything like the uniquely interpersonal reactive (full-scale) attitude of betrayal, and suggest that such a thinner account of trust is actually more useful in light of how they draw the trust/reliance distinction. Agential and non-agential accounts therefore seem to provide mutually exclusive, yet plausible, accounts of practices that we commonly refer to under the label of trust.

Is there perhaps a more circumscribed insight we can draw from our discussion? One interesting observation is this: Whereas non-agential accounts face some challenges when it comes to drawing practically relevant distinctions in *interpersonal* contexts (recall the Kant example: agential accounts do, after all, track something significant when they distinguish trust/reliance in terms of normative expectations etc.), agential accounts fall short when it come to offering practical guidance in context where we interact with *less-than-full* agents (it is, after all, very significant whether we merely rely on some technology or whether we deeply integrate it into our practical agency). Perhaps one tempting thought then is to search for some conceptual rendering that constitutes a 'middle ground' insofar as it includes sufficient 'normative resources' to allow us to distinguish normatively rich interpersonal trust from reliability and, at the same time, is sufficiently open-ended to account for non-agential settings (Nickel's account, discussed in §1.2 comes to mind—see my further discussion at the end of section 5 below).

## 6 Non-Ideal Theory: Concept Mismatch, Distrust, and Therapeutic Trust

My analysis up to this point has proceeded on the assumption that all reasoners within the community adopt either an agent-centered or a non-agential conception of trust and, moreover, that people use the concept in a way that matches their explanation of it. Put differently, I have only considered the difference of having our concept track one or the other understanding of trust-aptness, without assessing what may happen if we try to bring people to accept one or the other reading here and now, that is, in a setting where some people seem to adopt the former and some the latter, and moreover, it is not clear that individuals are consistent between their explanations and their usage of the concept.

But the most important worries that defenders of agent-centered accounts of trust have voiced against non-agential accounts, I want to suggest now, are worries that arise in *non-ideal circumstances* like the ones just described. Here is why: One of the agent-centered theorists' major worry I had mentioned earlier is that, through extending trust-aptness and popularizing this idea (e.g., in relation to AI systems), theorists could contribute to a suboptimal situation where, on the one hand, concept users employ the wider extension of trust proposed by non-agential accounts, yet,

on the other hand, they *retain* a number of implicit (positive, affective) inferences about trustworthiness /responsiveness that only make sense if our concept of trust is some version of R+AC. Ryan, for example, objects to the extension of trust to (non-agential) AI by suggesting that "one can say that [concept users] trust artefacts, such as AI, but this type of 'quasi trust' is actually *misplaced* trust. This type of misplaced trust has the potential to deceive individuals about the capacities of AI and obfuscate responsibility by AI companies." (2020, 251). Focusing on the connection between reactive attitudes and (moral) responsibility, Al makes a similar argument when he suggests that "continuing to apply the language of trust and trustworthiness to AI potentially gives the incorrect impression that developers and companies that employ AI either share responsibilities with the AI system or bear no responsibility at all when these systems fail." (2022, 11)

In the terminology I have just introduced, we would here be dealing with 'non-ideal' concept implementation or adoption as users are not fully complying with *either* an agent-centered or non-agential account's intension and extension. In line with the later, they would extend the language of trust to non-agents; but in line with the former, they retain the full gamut of emotional and accountability-tracking responses, including responses that are inapt and misleading when applied to AI technology and other non-agents. One upshot of trust under ideal theory where people knowingly adopt either account is that they will not anthropomorphize and make the kinds of mistakes to which Ryan, Al, and other point: under ideal conditions, reasoners will *either* attribute trust to non-agents in their thought *and* use a concept of trust whose intension is compatible with non-agential forms of trust (e.g. the unquestioning attitude account), *or* they will use an agent-centered account of trust *and* do not attribute trust to non-agents (except when doing so metaphorically).

To make sense of the theoretical problem of mismatches between people's attributions of trust and their associations of what follows from trust(aptness), it is useful to turn to Haslanger's distinction between manifest and operative social concepts.[18] According to Haslanger, one's manifest concept of x is "the concept [one] thought [one] was guided by and saw oneself as attempting to apply" (2006, 98), whereas one's *operative* concept is the one that "best captures the distinction that [one] in practice draw[s]." (2006, 98). We can illustrate this distinction with one of Haslanger's examples, namely the concept of 'parent': Haslanger and Saul (2006), 99) explains how, in her children's school, the manifest concept of 'parent', that is, the concept that teachers and parents take themselves to use in official communications, is most likely along the lines of 'direct biological progenitor'. In other words, this is the explanation of 'parent' that they would offer if asked about the meaning of the concept. By contrast, the operative concept of 'parent', that is, the definition that most closely tracks the actual use to which the concept is put, is different: it is the notion of 'primary caregiver', whether or not they are biologically related to the child.

[18] Haslanger famously also goes on to explain that we should sometimes aim for 'ameliorative' concepts/accounts of social kinds like gender or race. The connection between conceptual engineering and 'ameliorative' (conceptual) analysis is discussed in: (Haslanger, 2020a; Haslanger, 2020c)

Putting Haslanger's terminology to work for the issue of trust, the worry by critics of assessing AI in terms of trust/trustworthiness can be formulated like this: we should avoid scenarios where the conceptual extension implicit in a reasoner's use diverges from the properties that the reasoner takes themselves to apply in using the concept Why? Most obviously, we would end up misattributing qualities and properties to things that do not/cannot have them. Less obviously, we may more easily be influenced by those pushing a particular ideological or political agenda about who is responsible (and ultimately liable) in matters of advanced technology.

### 6.1 Assessing the Mismatch Worry

Though worries about extending the concept of trust into the domain of AI are frequently made, I think there are difficulties with this line of argument which have gone essentially unnoticed. My first concern is that it is not at all clear that we should accept as true the claim that a discrepancy between manifest and operative concept uses will arise for many competent concept users of trust if trust is applied to non-agents. In other words, the perceived confusion between 'manifest' and 'operative' concepts is often more imagined than real. Here is why: there are several intelligible renderings of the concept of trust that, if they *were* the concepts manifest with the reasoner, would avoid a manifest/operative mismatch. I think this objection is further supported by the empirical observation that those who think that AI agents can be assessed in terms of trust/trustworthiness are not more likely to attribute full moral responsibility to such systems (Malle & Ullmann, 2020).

Putting the accuracy of this claim to one side, suppose we did encounter a mismatch between operative and manifest concepts amongst a significant number of reasoners in a community of concept users. Does it really follow that we should try to restrict the extension that people have in their everyday use of the concept of trust to those for which it would be apt to apply them in accordance with what is people's manifest concept? The short answer is: it depends. This is because there are, after all, *two ways* of getting rid of a manifest/operative mismatch: one—as proposed by those critical of extending the concept of trust to AI systems—is to reduce the operative concept's reach (e.g. by educating people that AI cannot be trustworthy). The other one is to change the manifest concept (i.e. the explanation of the concept) that people have when doing so. Recall from the previous section that there really is no winner between agential and non-agential accounts of trust so long as people are aware of the explanation they are using. So unless it is somehow impossible to alter the manifest concept from something that conforms to the agential picture to the non-agential picture, there is no reason why this strategy of avoiding mismatch shouldn't be chosen.

### 6.2 The importance of non-agential distrust and 'therapeutic' dynamics

The strongest response to this line of argument, it seems to me, is this: even if people who talk of trustworthy AI adapted their manifest concepts to the non-agential account, it is unlikely that they would be able to rid themselves of the positive associations and attributions of responsibility and accountability (often subconscious

and implicit) that come with the attribution of trust and trustworthiness. It is very hard for people to avoid, when using the language/concept of trust, to make implicit inferences about trust-apt entities. And this may turn out to be deeply problematic for the case of AI technology because it renders ordinary concept users more pliant and gullible to the messaging of those pushing a particular ideological agenda, e.g., aimed a deflecting personal responsibility and corporate accountability.

My reply to this worry is that *even if* the use of the language of trust with the inappropriate implicit inference in AI entities is a real problem, it may not be altogether decisive. The reason is this: Considering X trust-apt is not only a necessary condition for judging it *trustworthy* (and therefore being disposed to be manipulated into positive affective/emotional states with regards to it): it is also a necessary condition for judging it *untrustworthy* and displaying an attitude of *distrust* towards it. Absent an assumption of trust-aptness, the only attitude one can take towards an artefact or other non-agents is to decide not to rely on it. And if the value of being able to distrust (and *communicate* distrust—see below), outweighs the risk of becoming gullible to proposals about trustworthiness, then we may, all things considered, favor extending trust-aptness in spite of these risks.

Though this is of course partly an empirical question about the downstream effects of concept use amongst reasoners that is hard to predict, I am inclined to believe that it matters more, here and now, that concept users are capable of leveraging the 'negative implicit associations' associated with the concept of distrust than the commensurate costs that come from implicit positive connotations that are potentially triggered when the concept of trust is used. The upshot of this argument is this: Conceptual engineers should aim to get people to operate with an operative concept that extends to non-agents like AI systems because it is essential that they be able to distrust, rather than merely not rely, on certain forms of technology. When technology is hard to understand and implemented by large corporate actors whose aims potentially diverge from those that are more or less forced to use them, then much is gained if our language/concept use allows for emotional distance, anger, and resentment towards these technologies. (Of course, there are two options here: where concept users already operate with a non-agential explanation as their manifest concept, 'distrust' will amount to a disposition to constantly monitor and question the relevant technology. By contrast, where concept users have a mismatch because 'distrust', following the agential model, is a failure of a commitment or an obligation, the more emotional register of resentment etc. will be at play.)

The important point is that the ability to distrust may be practically useful to have, irrespective of whether or not it is apt to experience the implicitly associated negative affective/emotional attitudes with these entities: it is one thing whether or not it is fitting to experience a particular affective/emotional response towards some entity; it is quite another whether or not having this response does or does not serve us well from a standpoint of practical agency.

It seems to me that there is a related argument here that concerns the phenomenon discussed under the label of 'therapeutic' (or 'proleptic') trust in the philosophical literature: The is idea is that, sometimes, extending the attitudes of trust to those who do not (yet) fully conform to the full-set of characteristics that would render trust warranted has beneficial consequences. The standard cases in the literature

concern interpersonal episodes where the trustor does not have (and rationally ought not to have) a fully fledged belief that the trustee will do what they are entrusted to do (Pettit, 1995, 199; Holton, 1994; Hieronymi, 2008; Frost-Arnold, 2014; Pace, 2021; Carter, 2022). In such instances one can trust the agent nonetheless and such trust is 'good trust' when it is aimed at bringing about the condition that will make the agent fully trustworthy.

My suggestion is that there can be a related analogue effect when it comes to bestowing trust on technological artefacts and other non-agents, whatever their actual status in terms of trust-aptness. The idea is this: By communicating that our reliance amounts to trust (in full knowledge that at least some concept users connect trust to reactive attitude and accountability practices), we make others aware that we believe that the relevant interaction ought to be assessed in terms of the more normative apparatus and expectations associated with trust. When others are operating with an agent-centered manifest account of trust, they will see this as an invitation to believe that normative expectations with regard to performance of the entrusted action are in order. This communicative process itself may have positive consequences.

Of course, this form of 'responsibilization' through therapeutic trust is not appropriate in just any given scenario: whether it is depends on the level of risk that trusting entails, as well as the degree of credence one can have for believing that the trust placed will not in fact be disappointed. In relation to technological artefacts and AI systems, we also need a sound hypothesis about how 'trust feedback' is likely to causally shape design processes. What is noteworthy here is that responsibilization need not be based on displaying *trust*. The therapeutic element could also be *distrust*. In this case, the trustor communicates a disposition to treat the relevant artefact or technology as capable of being either trusted or distrusted and the technology's designer responds to the social background norms according to which artefacts and systems of this kind are to be assessed as trustworthy or distrustworthy rather than merely being relied on.[19]

What is the connection between the phenomenon just described and the question of whether we should advance an agential or non-agential explanation of trust under conditions of non-ideal concept implementation? Suppose philosophers aimed to restrict 'trust proper' to human agents. If successful, it would mean that 'therapeutic' forms of trust-communication towards non-agents would more likely encounter the response that trusting them was 'a category mistake' to start with: "Don't talk about (dis)trust in relation to technology—all you could ever have done was rely on it!" This seems to deprive us of some useful ways of deploying the language of trust for practical purposes. However, there is also a reverse problem, so to speak, in case non-agential theorists get their way and all language users adopt a completely non-agential concept that lacks normative expectations: In this scenario, the 'therapeutic' element would lose some of its force because the fact that the object is trust-apt would no longer convey the

[19] (Chen, 2021, 1433) discusses the idea that the 'default' attitude to much AI technology should be one of distrust. The question of distrust in AI agents is all the more urgent given recent evidence (Krügel et al., 2022, 17) that laypeople seem to be more likely to place trust in untrustworthy AI agents than untrustworthy human agents.

'higher' level of normative expectations. One potential upshot of this may well be that, like I intimated in §4.6, there is some reason to favor 'moderate' non-agential accounts, like Nickel's (2022) that continue to attribute a significant role to normative expectations, but allow for trust in non-human agents by way of recognizing that such normative expectations can (and should) be directed at non-human agents and artefacts.

## 7 Conclusion

The purpose of this article has been to assess the prospect of explaining the concept of trust in line with agent-centered or non-agential accounts. Approaching this topic through the lens of conceptual engineering has offered at least two advantages: First, the conceptual engineering framework helpfully requires us to describe explicitly the various functional characteristics that our concept of trust should have and to evaluate different accounts against such a functional standard.

Second, in laying out the implications of agential vs. non-agential accounts of trust in relation to their ability to distinguish trust from reliance, propose an explanatorily useful account, illuminate the affective/emotional aspects of our practice, and provide an explanation that highlights the connection between trust and distrust, I hope to have shown that the relevant choice is one between two conceptual renderings that, though mutually exclusive, each offer a plausible and attractive understanding of trust, and one that chimes very well with at least some parts of our central practices in this domain. As a result, it is not entirely clear whether we can reach any clear verdict about which rendering of trust should prevail.

To conclusively settle this matter, we would need to have greater knowledge about how some difficult-to-predict dynamics of concept use and their relations to implicit inferences (positive and negative) would play out if concept users altered their manifest concepts. My own attempt, in section 5, to offer some arguments about how we should think through the potential consequences of advocating and popularizing agential and non-agential accounts under current circumstances constituted an initial step in this direction, if only a preliminary one: it is not obvious, as some critics of non-agential accounts have suggested, that a broader account of trust is likely to confuse users and manipulate them into attributing trust where they shouldn't. And even if there is such a risk, there are important conceptual resources that extending trust to non-agents makes available, perhaps most notably the possibility that, when it comes to negative attitudes, we can move beyond merely not relying on non-agents but also *distrust* them.

As technology advances rapidly and comes to shape our practical agency in ever more comprehensive ways, it is likely that pressure to assess non-human agents in terms of our received concepts will increase. This is likely true not only for whether or not human trust can be aptly placed in non-human agents and artefacts (the topic of the literature to which this article contributed), but also for the topic of whether or not we should assess non-human agents as engaging in practices of trust. For this reason, reflection on how we should (re)define our concepts becomes more urgent. Technologies must be fit for purpose— but so too must our concepts.

**Acknowledgements** I would like to thank Markus Kneer, Herman Veluwenkamp, and Daniel Viehoff for discussion and participants of workshops on trust at the University of Zurich (Digital Society Initiative) and on conceptual engineering at the University of Delft.

**Authors' contributions** JV sole author

**Funding** No Funding received

**Data availability** Not applicable

## Declarations

**Ethics approval and consent to participate** Not applicable

**Consent for publication** Not applicable

**Competing interests** No competing interests to be declared.



## References

Al, P. (2022). (E)-Trust and Its Function: Why We Shouldn't Apply Trust and Trustworthiness to Human–AI Relations. *Journal of Applied Philosophy*. https://doi.org/10.1111/japp.12613

Baier, A. (1986). Trust and Antitrust. *Ethics, 96*(2), 231–260.

Bieber, F., & Viehoff, J. (2023). A Paradigm-Based Explanation of Trust. *Synthese, 201*(1), 2. https://doi.org/10.1007/s11229-022-03993-4

Brennan, J. (2021). Trust as a Test for Unethical Persuasive Design. *Philosophy & Technology, 34*(4), 767–783. https://doi.org/10.1007/s13347-020-00431-6

Bryson, Joanna. 2018. "AI & Global Governance: No One Should Trust AI." United Nations University. https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html.

Budnik, C. (2018). Trust, Reliance, and Democracy. *International Journal of Philosophical Studies, 26*(2), 221–239. https://doi.org/10.1080/09672559.2018.1450082

Buechner, J., & Tavani, H. T. (2011). Trust and Multi-Agent Systems: Applying the 'Diffuse, Default Model' of Trust to Experiments Involving Artificial Agents. *Ethics and Information Technology, 13*(1), 39–51. https://doi.org/10.1007/s10676-010-9249-z

Burgess, A., Cappelen, H., & Plunkett, D. (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Capellen, Herman, and David Plunkett. 2020. "A Guided Tour of Conceptual Engineering and Conceptual Ethics." In *Conceptual Engineering and Conceptual Ethics*, by Plunkett, David and Capellen, Herman, 230–260. Oxford University Press. https://doi.org/10.1093/oso/9780198801856.003.0012.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering* (First ed.). Oxford University Press.

Carter, J. A. (2022). Therapeutic Trust. *Philosophical Psychology, 0*(0), 1–24. https://doi.org/10.1080/09515089.2022.2058925

Chen, M. (2021). Trust and Trust-Engineering in Artificial Intelligence Research: Theory and Praxis. *Philosophy & Technology, 34*(4), 1429–1447. https://doi.org/10.1007/s13347-021-00465-4
Coeckelbergh, M. (2012). Can We Trust Robots? *Ethics and Information Technology, 14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1
D'Cruz, J. (2020). Trust and Distrust. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 41–51). Routledge.
DeCamp, M., & Tilburt, J. C. (2019). Why We Cannot Trust Artificial Intelligence in Medicine. *The Lancet Digital Health, 1*(8), e390. https://doi.org/10.1016/S2589-7500(19)30197-9
Faulkner, P. (2007). On Telling and Trusting. *Mind, 116*(464), 875–902. https://doi.org/10.1093/mind/fzm875
Faulkner, P. (2011). *Knowledge on Trust*. Oxford University Press.
Ferrario, A., Loi, M., & Viganò, E. (2020). In AI We Trust Incrementally: A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology, 33*(3), 523–539. https://doi.org/10.1007/s13347-019-00378-3
Ferrario, A., Loi, M., & Viganò, E. (2021). Trust Does Not Need to Be Human: It Is Possible to Trust Medical AI. *Journal of Medical Ethics, 47*(6), 437–438. https://doi.org/10.1136/medethics-2020-106922
Fricker, M. (2019). Forgiveness—An Ordered Pluralism. *Australasian Philosophical Review, 3*(3), 241–260. https://doi.org/10.1080/24740500.2020.1859230
Fricker, Miranda. 2023. "Diagnosing Institutionalized 'Distrustworthiness.'" *The Philosophical Quarterly*, March, pqad031. https://doi.org/10.1093/pq/pqad031.
Frost-Arnold, K. (2014). The Cognitive Attitude of Rational Trust. *Synthese, 191*(9), 1957–1974. https://doi.org/10.1007/s11229-012-0151-6
Goldberg, S. C. (2020). Trust and Reliance. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge.
Gordon, E. C. (2022). When Monitoring Facilitates Trust. *Ethical Theory and Moral Practice, 25*(4), 557–571. https://doi.org/10.1007/s10677-022-10286-9
Grodzinsky, F., Miller, K., & Wolf, M. J. (2020). Trust in Artificial Agents. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge.
Hardin, R. (1996). Trustworthiness. *Ethics, 107*(1), 26–42. https://doi.org/10.1086/233695
Haslanger, S. (2012). *Resisting Reality: Social Construction And Social Critique*. Oxford University Press.
Haslanger, S. (2020a). How Not to Change the Subject. In T. Marques & A. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability* (p. 235). Oxford University Pres.
Haslanger, S. (2020b). Going On, Not in the Same Way. In *Conceptual Engineering and Conceptual Ethics, by Sally Haslanger* (pp. 230–260). Oxford University Press. https://doi.org/10.1093/oso/9780198801856.003.0012
Haslanger, S. (2020c). Going On, Not in the Same Way. In H. Capellen & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp. 230–260). Oxford University Press.
Haslanger, S., & Saul, J. (2006). Philosophical Analysis and Social Kinds. *Proceedings of the Aristotelian Society, Supplementary Volumes, 80*, 89–143.
Hatherley, J. J. (2020). Limits of Trust in Medical AI. *Journal of Medical Ethics, 46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935
Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs, 48*(1), 1–20. https://doi.org/10.1111/nous.12000
Hawley, K. (2017). Trustworthy Groups and Organizations. In P. Faulkner & T. W. Simpson (Eds.), *The Philosophy of Trust*. Oxford University Pres.
Hieronymi, P. (2008). The Reasons of Trust. *Australasian Journal of Philosophy, 86*(2), 213–236. https://doi.org/10.1080/00048400801886496
Himmelreich, J., & Köhler, S. (2022). Responsible AI Through Conceptual Engineering. *Philosophy & Technology, 35*(3), 60. https://doi.org/10.1007/s13347-022-00542-2
Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy, 72*(1), 63–76. https://doi.org/10.1080/00048409412345881
International Astronomical Union. (2006). *IAU 2006 General Assembly Resolution*. Press Release https://www.iau.org/news/pressreleases/detail/iau0603/
Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual Engineering: A Road Map to Practice. *Philosophy Compass, 17*(10), e12879. https://doi.org/10.1111/phc3.12879

Jones, K. (1996). Trust as an Affective Attitude. *Ethics, 107*(1), 4–25.
Jones, K. (2012). Trustworthiness. *Ethics, 123*(1), 61–85.
Kneer, M. (2021). Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. *Cognitive Science, 45*(10), e13032. https://doi.org/10.1111/cogs.13032
Koch, S. (2018). *The Externalist Challenge to Conceptual Engineering*. *Synthese*.
Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *Philosophy & Technology, 35*(1), 17. https://doi.org/10.1007/s13347-022-00511-9
Malle, Betram, and Daniel Ullmann. 2020. "A Multidimensional Conception and Measure of Human-Robot Trust." In *Trust in Human-Robot Interaction*, by Chang S. Nam and Joseph B. Lyons. https://doi.org/10.1016/B978-0-12-819472-0.00001-0.
Margolis, E., & Laurence, S. (2023). Concepts. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
McGeer, V. (2008). Trust, Hope and Empowerment. *Australasian Journal of Philosophy, 86*(2), 237–254. https://doi.org/10.1080/00048400801886413
McGeer, V., & Pettit, P. (2017). The Empowering Theory of Trust. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 14–34). Oxford University Press.
McLeod, C. (2002). *Self-Trust and Reproductive Autonomy*. MIT Press.
Nado, J. (2021). Classification Procedures as the Targets of Conceptual Engineering. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12843
Nguyen, C. T. (2022). "Trust as an Unquestioning Attitude." In *Oxford Studies in Epistemology* Vol. 7, edited by Tamar Szabó Gendler, John Hawthorne, and Julianne Chung. Oxford University Press. https://doi.org/10.1093/oso/9780192868978.003.0007
Nickel, P. J. (2007). Trust and Obligation-Ascription. *Ethical Theory and Moral Practice, 10*(3), 309–319. https://doi.org/10.1007/s10677-007-9069-3
Nickel, P. J. (2017). Being Pragmatic about Trust. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust*. Oxford University Pres.
Nickel, P. J. (2022). Trust in Medical Artificial Intelligence: A Discretionary Account. *Ethics and Information Technology, 24*(1), 7. https://doi.org/10.1007/s10676-022-09630-5
O'Neil, C. (2012). Lying, Trust, and Gratitude. *Philosophy & Public Affairs, 40*(4), 301–333. https://doi.org/10.1111/papa.12003
Owens, D. (2017). Trusting a Promise and Other Things. In *The Philosophy of Trust, by Paul Faulkner and Thomas Simpson*. Oxford University Press.
Pace, M. (2021). Trusting in Order to Inspire Trustworthiness. *Synthese, 198*(12), 11897–11923. https://doi.org/10.1007/s11229-020-02840-8
Pettit, P. (1995). The Cunning of Trust. *Philosohpy and Public Affairs, 24*, 202–225.
Rawls, J. (2001). *Justice as Fairness : A Restatement*. Harvard University Press.
Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y
Simion, M., & Kelp, C. (2020). Conceptual Innovation, Function First. *Noûs, 54*(4), 985–1002. https://doi.org/10.1111/nous.12302
Simpson, T. (2012). What Is Trust? *Pacific Philosophical Quarterly, 93*(4), 550–569. https://doi.org/10.1111/j.1468-0114.2012.01438.x
Soter, S. (2006). What Is a Planet? *The Astronomical Journal, 132*(December), 2513–2519. https://doi.org/10.1086/508861
Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2022). Intentional Machines: A Defence of Trust in Medical Artificial Intelligence. *Bioethics, 36*(2), 154–161. https://doi.org/10.1111/bioe.12891
Stuart, M. T., & Kneer, M. (2021). Playing the Blame Game with Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI'21 Companion), March 8-11, 2021, Boulder, CO* https://www.academia.edu/45077358/Playing_the_Blame_Game_with_Robots
Taddeo, M. (2010). Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust. *Minds and Machines, 20*(2), 243–257. https://doi.org/10.1007/s11023-010-9201-3
Tallant, J. (2019). You Can Trust the Ladder, But You Shouldn't. *Theoria, 85*(2), 102–118. https://doi.org/10.1111/theo.12177
Tavani, H. T. (2015). Levels of Trust in the Context of Machine Ethics. *Philosophy & Technology, 28*(1), 75–90. https://doi.org/10.1007/s13347-014-0165-8

Thomasson, Amie. 2021. "Conceptual Engineering: When Do We Need It? How Can We Do It?" *Inquiry* 0 (0): 1–26. https://doi.org/10.1080/0020174X.2021.2000118.

Ullman, D., & Malle, B. F. (2018). What Does It Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 263–64. HRI '18*. Association for Computing Machinery. https://doi.org/10.1145/3173386.3176991

Veluwenkamp, H., Capasso, M., Maas, J., & Marin, L. (2022). Technology as Driver for Morally Motivated Conceptual Engineering. *Philosophy & Technology, 35*(3), 71. https://doi.org/10.1007/s13347-022-00565-9

Veluwenkamp, H., & van den Hoven, J. (2023). Design for Values and Conceptual Engineering. *Ethics and Information Technology, 25*(1), 2. https://doi.org/10.1007/s10676-022-09675-6

Walker, M. U. (2006). *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge University Press.

Weiskopf, D. A. (2009). The Plurality of Concepts. *Synthese, 169*(1), 145–173. https://doi.org/10.1007/s11229-008-9340-8