TRUST, VULNERABILITY, AND MONITORING

I. INTRODUCTION

Here are two perennial questions in the philosophy of trust, both of which concern the relationship between trust and risk:

> *Vulnerability question*: In what sense does trusting essentially involve subjecting oneself to risk of betrayal?
>
> *Monitoring question*: In what sense is monitoring for risks of betrayal incompatible with trusting?

These questions have traditionally been pursued independently from one another.[1] It will be shown that they are much more closely connected than has been appreciated. The central objective will be to demonstrate how a performance-normative framework can be used to answer both the Vulnerability Question and the Monitoring Question in a principled way, one that reveals a deep connection between not just the questions themselves, but also between the concepts of vulnerability, monitoring, and *de minimis* risk.

II. TRUST AND VULNERABILITY TO BETRAYAL

The very idea that trusting constitutively involves subjecting oneself to the risk that one's trust is betrayed is platitudinous in the philosophy of trust.[2] But what counts as 'subjecting oneself' to risk of betrayal? Getting

[1] For discussions of the relationship between trust and monitoring, see, e.g., Hieronymi ("The Reasons of Trust," *Australasian Journal of Philosophy* 86, no. 2 (2008): 213–36) and McMyler (*Testimony, Trust, and Authority* (OUP USA, 2011)) and Wanderer and Townsend ("Is It Rational to Trust?" *Philosophy Compass* 8, no. 1 (2013): 1–14). For some representative discussions of trust's relationship to vulnerability, see e.g., Nickel and Vaesen ("Risk and Trust," in *Handbook of Risk Theory*, ed. Sabine Roeser et al. (Springer, 2012), 861–62). Cf., Pettit ("The Cunning of Trust," *Philosophy and Public Affairs* 24, no. 3 (1995): 208).

[2] For various expressions of this idea, see, along with Hardin ("The Street-Level Epistemology of Trust," *Analyse & Kritik* 14, no. 2 (1992): 152–76), e.g., Baier ("Trust and Antitrust," *Ethics* 96, no. 2 (1986): 244), McLeod ("Trust," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2020 (Metaphysics Research Lab, Stanford University, 2020), sec. 1), Nickel and Vaesen ("Risk and Trust," 861–62), Becker ("Trust as Noncognitive Security about Motives," *Ethics* 107, no. 1 (1996): 45, 49), Dormandy ("Exploitative Epistemic Trust," in *Trust in Epistemology*, ed. Katherine Dormandy, 2020, 241–42), Kirton ("Matters of Trust as Matters of Attachment Security," *International Journal of Philosophical Studies*, forthcoming, 1–20), O'Neil ("Betraying Trust," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas W. Simpson (Oxford, UK: Oxford University Press, 2017), 70–72), and Hinchman ("On the Risks of Resting Assured: An Assurance Theory of Trust," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas W. Simpson (Oxford: Oxford University Press, 2017)). Cf., Pettit ("The Cunning of Trust," 208).

this right is important to understanding the nature of trust and what is distinctive about it.

One tempting starting point – widespread in the social and behavioural sciences[3] – is to begin with the role that trust plays in facilitating cooperation between parties with competing interests. And here a common view maintains that trust functions as a strategy to mitigate, without entirely eliminating, uncertainty.[4]

This way of thinking suggests a natural, even if imperfect[5], contrast between trusting someone X to ϕ (as entrusted) with *knowing* that X will do so – one that invites us to link trust-relevant vulnerability to betrayal with (some non-negligible degree of) *ignorance* about whether trustee will come through.[6]

Unfortunately, this kind of a starting point only gets us so far. It invites us to ask – what *kind* of ignorance suffices here? On the one hand, one might be ignorant that a trustee X will come through simply because there is some *actual* risk, *R*, (above some threshold) to X's coming through, and *regardless* of whether S perceives this to be the case. This is called *objective risk*[7]; it is objective because its status as a risk doesn't non-trivially depend on its being perceived as such. For example, an impending storm presents a risk that you will not be able to finish painting the house as entrusted, even if you are in denial – or misinformed about

[3] See, e.g., Krishnan et al. ("When Does Trust Matter to Alliance Performance?" *The Academy of Management Journal* 49, no. 5 (2006): 894–917), Waston and Moran (*Trust, Risk, and Uncertainty* (Palgrave-Macmillan, 2005)), Beck (*Risk Society: Towards a New Modernity*, vol. 17 (sage, 1992)).

[4] As Frederiksen ("Trust in the Face of Uncertainty: A Qualitative Study of Intersubjective Trust and Risk," *International Review of Sociology* 24 (2014): 130–44) puts it, 'Contemporary trust research regards trust as a way of dealing with uncertainty and risk. Predominantly, it suggests that trust reduces uncertainty by means of risk assessment and rational calculation.'

[5] The Cartesian position that knowledge entails subjective certainty no longer enjoys much popularity in mainstream epistemology. Though cf., Beddor ("New Work For Certainty," *Philosophers' Imprint* 20, no. 8 (2020)) for discussion.

[6] The idea that knowledge obviates the need for trust is broadly analogous to the thought, due to Plato, that knowledge obviates the need for inquiry. In the *Meno*, Plato maintains that one 'cannot inquire about what he knows, because he knows it, and in that case is in no need of inquiry (Plato, *Plato's Meno*, ed. Richard Stanley Bluck (385BC; repr., Cambridge University Press, 2011), sec. 80.e). The idea under consideration proceeds by a similar reasoning: 'one cannot trust another to do what he knows he will do, because he knows he will do it, an in that case there is no need for trust.' A contemporary variation on this idea is found in the sociology of George Simmel, who explicitly contrasts trusting with knowing (see, e.g., Wolff *The Sociology of Georg Simmel* (Glencoe, Ill: Free Press, 1950)).

[7] For discussion, see Hansson ("Risk," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2018 (Metaphysics Research Lab, Stanford University, 2018)).

– the weather forecast. On the other hand, one might be ignorant that X will come through simply because one *perceives* there to be some risk (even if, objectively, there is not). *Perceived risk* is such that its status as a risk *does* (non-trivially) depend on its being perceived as such.[8] For example, the *perceived risk* that 5G towers increases the spread of Covid is such that its status as a risk is entirely dependent upon its (mistaken) perception as such.[9]

The distinction between objective and perceived risks maps naturally on to two different ways of answering the Vulnerability Question. According to a simple perceived-risk account of trust-relevant vulnerability to betrayal, trust essentially involves subjecting oneself to *perceived* risk of betrayal, though not to *objective* risk of betrayal.

I will argue in §II.a that the simple perceived risk account is untenable. Trusting essentially involves subjecting yourself to at least some objective risk of betrayal. But this raises a question: what is the right way to characterise the kind of objective risk to which, by trusting, one essentially subjects herself? In §II.b I consider and reject two answers: as (i) the product of the estimated objective probability of betrayal multiplied by the disvalue of betrayal (i.e., as *risk expectation value*); and as (ii) the objective (frequentist) probability of betrayal alone, above some specified threshold. What the defects in these accounts reveal is the need for a *normative* objective account – framed in terms of *de minimis* risk – which is what I'll go on to propose and defend.

II.a. A simple perceived risk account

One initial – but ultimately misguided – line of argument against a simple perceived risk account of trust-relevant vulnerability to betrayal holds that there is a tension between (i) the presumed explicit, conscious awareness involved in risk perception; and (ii) the unconscious or tacit character of (at least some kinds of) trusting. Trust can certainly be unconscious or tacit.[10] And it *seems* plausible on first blush that risk perception is not. For

[8] See, e.g., Sjoberg ("Explaining Risk Perception," *An Evaluation of the Psychometric Paradigm in Risk Perception Research* 10, no. 2 (2004): 665–12) and Slovic ("Perception of Risk," *Science* 236, no. 4799 (1987): 280–85).

[9] For discussion of this perceived risk, and the extent of its uptake on social media, see Ahmed et al. ("COVID-19 and the 5g Conspiracy Theory: Social Network Analysis of Twitter Data," *Journal of Medical Internet Research* 22, no. 5 (2020): 19–45).

[10] For some empirical discussion on the ubiquity of tacit trust, see, e.g., Lagerspetz (*Trust: The Tacit Demand*, vol. 1 (Springer Science & Business Media, 1998)), Burns ("Explicit and Implicit Trust Within Safety Culture," *Risk Analysis* 26, no. 5 (2006): 1139–50) and Guo et al.

example, it is a hallmark of the 'Risk Society' research programme[11] that our perceptions of risk are often given expression through affect such as fear and anxiety.[12]

But the tension here is only apparent. The countenancing of implicit trust is problematic for the perceived risk account only if risk perception of the sort that is essential to trust can't *itself* be unconscious or tacit. But the empirical evidence – especially over the past several decades[13] – on unconscious bias and risk perception has established, uncontroversially, that *even if* some risk perception is accompanied with conscious awareness (i.e., some combination of occurrent beliefs plus affect) a significant extent of our risk perception takes place below the surface of conscious awareness. (Compare: our cognitive biases are often *unconscious* biases, and at least some of these biases consist in perceptions of risk[14]).[15]

What this means is just that *if* the perceived risk account of trust relevant vulnerability to betrayal is problematic, it isn't going to be so because of any 'mismatch' between the implicit character of (some) trust and the alleged conscious character of risk perception; just as trust itself can be deliberative or implicit, so can our perceptions of risks to its being betrayed. There is, however, a much more serious problem that faces the perceived risk account of trust-relevant vulnerability to betrayal. Consider the following case:

> SUNRISE: Having read some fringe QAnon conspiracy theories on a Reddit subthread, you come to think your friend is among a select group of people who decides how and when the sun rises, by manipulating the earth's orbit and rotation.

("From Ratings to Trust: An Empirical Study of Implicit Trust in Recommender Systems," in *Proceedings of the 29th Annual Acm Symposium on Applied Computing*, 2014, 248–53).

[11] Beck *Risk Society*; Goddens *The Consequences of Modernity* (John Wiley & Sons, 2013).

[12] In this line of thinking, Bauman (*Liquid Fear* (John Wiley & Sons, 2013)) describes our modern high-tech predicament, characterised by new technologies and dangers, as pervaded by a 'derivative fear' namely 'the sentiment of being susceptible to danger: a feeling of insecurity and vulnerability.'

[13] See, e.g., Sjoberg ("The Methodology of Risk Perception Research," *Quality and Quantity* 34, no. 4 (2000): 407–18) and Slovic ("Risk Perception," in *Carcinogen Risk Assessment* (Springer, 1988), 171–81).

[14] One classic example here is 'shooter bias' (e.g., Unkelbach et al. "The Turban Effect: The Influence of Muslim Headgear and Induced Affect on Aggressive Responses in the Shooter Bias Paradigm," *Journal of Experimental Social Psychology* 44, no. 5 (2008): 1409–13).

[15] For some representative discussions of unconscious or implicit bias, which include some perceptions of risk, see, e.g., Saul ("Scepticism and Implicit Bias," *Disputatio* 5, no. 37 (2013): 243–63) Holroyd et al. ("What Is Implicit Bias?" *Philosophy Compass* 12, no. 10 (2017): e12437).

> Afraid the group might trigger an event that would shroud your hemisphere in permanent darkness (something you believe you friend has final control over), you say "Can I trust you not to prevent the sun from rising?" Your friend (though finding this request strange) says they can surely oblige, simply because they knew that betrayal here would be impossible.

Question: Did you really *trust* your friend to not prevent the sun from rising, or did you merely *think* you did? There are two good reasons to think you merely *thought* you did. The first appeals to a very weak attribution principle according to which S trusts X with ϕ only if either (i) X is in a position to have ϕ-ing attributed to her; or X is in a position to have not-ϕ-ing attributed to her. This principle is implied the platitude that trustees incur any commitments at all *vis-à-vis* what they are entrusted to do, commitments they may uphold or not depending on what the trustee does. Granted, one could reject this attribution principle, but only on pain of then losing a grip on what distinguishes trustees from those (e.g., mere sympathisers with the trustor, bystanders, etc.) who incur no commitments to the trustor, *vis-à-vis* ϕ-ing, one way or another. But, crucially, from this attribution principle it follows straightforwardly that you didn't really trust your friend in SUNRISE, even if you thought you did.

A second reason for doubting that genuine trust is present when you think you're trusting but subjecting yourself to *merely* perceived risk (i.e., as is the case in SUNRISE) is closely related to the first. Just consider the tight relationship between trust and reactive attitudes such as gratitude and blame. A common view in the philosophy of trust is that gratitude would be an appropriate or 'fitting' reactive attitude to a trustee's coming through, as blame would be to betrayal.[16] But, as this line of thought goes, gratitude would be *clearly* misplaced, if directed by your friend to *you*, when the sun then goes on to rise the next day as expected.

These points suggest that there is an intractable kind of problem with the simple perceived risk account. If trusting essentially involves subjecting oneself to *merely* perceived risk of betrayal, then there will be pressure to rule-in cases like SUNRISE as cases of genuine trust. But we have good independent grounds for thinking that SUNRISE is not a case of genuine trust; at least, this is the case given the very weak assumptions that trust

[16] See, e.g., O'Neil ("Lying, Trust, and Gratitude," *Philosophy & Public Affairs* 40, no. 4 (2012): 301–33), Domenicucci and Holton ("Trust as a Two-Place Relation," *The Philosophy of Trust*, 2017, 149–60), and D'Cruz ("Trust, Trustworthiness, and the Moral Consequence of Consistency," *Journal of the American Philosophical Association* 1, no. 3 (2015): 467–84).

involves (i) the incurring of normative commitments by the trustee *vis-à-vis* what she is entrusted with; and (ii) the presumed connection with the reactive attitudes it is taken to have.

The proponent of a simple perceived risk account of trust-relevant vulnerability to betrayal might press back by digging in the heels and defending an *immunity from error* thesis along the following lines: if one believes one is trusting X to ϕ, then one is trusting X to ϕ. If this immunity from error thesis is correct, then I am wrong to claim that (given the lack of any objective risk) I merely *think* I am trusting you not to fiddle with the earth's orbit. Whether I'm trusting you with something is, as this line of thought goes, not something I could be mistaken *about*: it is guaranteed that I am trusting you with X if I take myself to be trusting you with X. Trusting is in this respect akin to the kinds of mental states (i.e., perhaps like 'being in pain' or 'being confused') that have the property of being such that if one believes one is in that state, then one is in that state.[17]

But such an immunity from error thesis is false in the case of trust for two main reasons, and the perceived risk account of trust-relevant vulnerability to betrayal therefore can't press back against the objections raised by relying on it. The first reason has to to do with the the fact that we can and often are mistaken about reliance facts. Reliance is necessary even though not sufficient for trust.[18] However, I am not infallible about whether I am relying on you for X; for one thing, I might forget I am relying on you to repay a debt. Or, I might forget that you've already repaid a debt and so believe mistakenly that I am *still* relying on you to repay it. But since I can mistake reliance facts, as the thought goes, my thoughts about whether I trust can't be self-guaranteeing.

Secondly, and following here Santiago Echeverri,[19] one standard way to test whether a belief that you are in some state guarantees its own truth is to ask whether it would be either incoherent or irrational for one to *question* whether one is in that state.[20] But for any case where we have trusted someone X to ϕ, we can coherently question whether we have done so.

[17] For discussion, see, e.g., Shoemaker ("Moore's Paradox and Self-Knowledge," *Philosophical Studies* 77, no. 2-3 (1995): 211–28) and Burge ("Reason and the First Person," *Knowing Our Own Minds*, 1998, 243–70).

[18] For discussion, see Carter and Simion ("The Ethics and Epistemology of Trust," *Internet Encyclopedia of Philosophy*, 2020).

[19] "Guarantee and Reflexivity," *Journal of Philosophy* 117, no. 9 (2020): 473–500.

[20] For example, it might be incoherent or irrational to ask whether you are a thinking thing, or (to use an example from Kaplan ("On the Logic of Demonstratives," *Journal of Philosophical Logic* 8, no. 1 (1979): 81–98) involving indexicals to ask whether it is true that "I am here now?"

This is thus another reason why it is a mistake to attempt to revive the perceived risk account by latching on to the idea that it's impossible to think you are trusting someone with something when you're not. All this points to is the beginnings of an answer to the Vulnerability Question. That question asks: In what sense does trusting essentially involve subjecting oneself to risk of betrayal? Our working answer is now: not *merely* to perceived risks of betrayal. Let's continue to refine this answer.

II.b. Towards an objective risk account

Let's explore now the idea that necessary to trusting is subjecting yourself to at least some non-negligible risk to betrayal whose status *as* a risk to betrayal doesn't (non-trivially) depend on its being perceived as such. A natural first-pass at refining this idea maintains the following: trust essentially involves subjecting oneself to risk of betrayal *beyond some objective risk 'threshold'.*

As is common in risk analysis[21], an (objective) risk threshold is set as (above or below) some specified risk *expectation value*, which is calculated as the product of (i) objective (or frequentist) probability of the risk event obtaining; and (ii) its severity (i.e., degree of harm of the risk event's obtaining). For example, the risk expectation value of a low-probability risk with significant severity were it to obtain might be very similar to the risk expectation value of a much higher-probability but less severe risk.

A qualification here needs some care. One might ask "Since we must inevitably use our own evidence to work out what the risk expectation value is for a given risk, and different people have different evidence that they will be relying on to make such an assessment, then doesn't the notion of 'objective' risk – understood as above a risk expectation value threshold – just collapse into a perceived risk account?" The answer, importantly, is 'no.' When we try to determine a given risk expectation value, we inevitably make a subjective *assessment* of the objective probability of the risk event obtaining as well as of the objective disvalue. But – and this is a crucial point of difference between the notions of objective risk and perceived risk – on the latter account, what the risk facts are do not depend on our *estimates.* In characterising risk expectation value, we are attempting to characterise something that is *mind-independent.* Perceived risks by contrast depend (non-trivially) on their being perceived as such.

[21] See, e.g., Hansson ("Risk"; Sven Ove Hansson, "Philosophical Perspectives on Risk," *Techné: Research in Philosophy and Technology* 8, no. 1 (2004): 10–35) for discussion.

Bearing these qualifications in mind, appealing to objective risk expectation value (the product of the objective probability of the risk event obtaining multiplied by its severity) would be an obvious way by which one might try to assess *risk of betrayal*, simpliciter. However, appealing to objective risk expectation value it is ultimately not a promising way to think about *trust-relevant* vulnerability to betrayal, viz., as vulnerability expressed in terms of risk expectation value threshold. The problem is not the objective frequentist interpretaion of probability at issue[22], but rather, what happens when we adjust (significantly) the expected disvalue. To see the problem, consider the following simple pair of cases:

> BABYSITTER: *A* trusts *B* to responsibly babysit their only child, *C*, for the weekend; assume the objective probability of betrayal is .001 and would generate 100,000 units of disvalue.
>
> PENCIL: *A* trusts *B* to use *A*'s pencil and return it; assume the objective probability of betrayal is .1 and betrayal would generate .001 unit of disvalue.

Both BABYSITTER and PENCIL are paradigmatic cases of trust, though the the risk expectation value products are dramatically different: in BABYSITTER, .001 x 100,000 = a risk expectation value of 100. In PENCIL, .1 x .0001 = a risk expectation value of .00001, which will – and here is the worry – end up being lower than any kind of plausible threshold we might appeal to in order to distinguish cases of genuine trust from cases where there is effectively no objective risk of betrayal. What's more, cases like PENCIL become even more difficult for the kind of proposal under consideration when we lower even further the disvalue of betrayal (e.g., to .00000001 disvalue).[23] What cases like PENCIL seem to suggest, then, is that if we want to characterise the kind of risk that trust essentially involves subjecting oneself to in terms of objective rather than merely perceived risk, we might do better to simply control for the severity of betrayal and then characterise the relevant risk threshold solely in terms of the objective probability of betrayal. Then, presumably, PENCIL will be

[22] It is worth noting that risk expectation value, while naturally allied to a probabilistic gloss, isn't necessarily tied to one. For a modal approach to risk expectation value, see Pritchard ("Risk," *Metaphilosophy* 46, no. 3 (2015): 436–61).

[23] Another kind of case that serves to capture this kind of problem will simply shift the value of what is entrusted to near zero, where the shift takes place after trust is placed. For example, I may loan you my gold pen (my only valuable possession) so you can impress a client. I trust that you'll return it. In the meantime, a goldmine might be discovered that saturates the market and sends its value to ~£0. This fact doesn't undermine my having trusted, and continuing to trust, you to return the gold pen. Thanks to [OMITTED] for discussion of this kind of case.

above the relevant risk threshold, given that the probability is .1 (10%).

Continuing with this idea, suppose we were to set the threshold as .05 (5%). This move will get PENCIL right; and since the probability that the sun won't rise is vanishingly low, there is no pressure to rule in SUNRISE. However, the cost with this kind of a move is that we can then no longer deal with cases like BABYSITTER. After all, when stakes are high (i.e., when the disvalue of betrayal is suitably high), it seems we might, and very often do, trust one even when the objective risk is very low – i.e., 1 in 1,000 (.10%) as in the case of BABYSITTER. Cases like BABYSITTER are not aberrations: many cases of trust (e.g., with loved ones' lives and welfare) have a structure whereby something of high value is entrusted to someone very reliable, precisely *because* they are very reliable, and are accordingly very unlikely to betray the trust.

One might try to deal with the above by simply setting the objective probability even lower. On such a view, trust essentially involves subjecting oneself to betrayal in the sense that the objective probability of betrayal must be, e.g., at least 0.000001 (0.0001%, i.e., 1 in a million). Since there is probably at *least* a 1 in a million chance the the hero in BABYSITTER brings about a disaster, this tweak seems to put the threshold on the right side of BABYSITTER. But the cost of setting the threshold *this* low is that you then invite an entirely different problem, which brings us back to cases in the vicinity of SUNRISE.[24]

II.c. A performance-normative account

Here is where we've got to. Trust essentially involves subjecting oneself to risk of betrayal in a sense that: (i) cannot be captured exclusively with reference to perceived risk of betrayal, because trust requires at least some objective risk betrayal; however, (ii) the threshold of objective risk above which one by trusting must essentially subject herself isn't something we can plausibly capture satisfactorily in terms of either (a) objective risk expectation value; or (b) the objective probability of betrayal alone; (iii) neither (a) nor (b) could handle all three of our examples cases together; and so (iv) *whatever* level of objective risk beyond which by trusting we thereby subject ourselves accordingly needs to be characterised in some other way.

[24] After all, once the threshold is set this low, then it will be difficult to explain why we should rule out (as we should) cases as being cases of genuine trust where the weak attribution principle isn't satisfied, and where reactive attitudes toward the would-be trustee would be misplaced.

The way forward, I want to suggest, is to pursue the idea that the relevant threshold of objective risk to which by trusting one essentially subjects herself is fixed by neither (i) risk expectation value nor by (ii) simple objective probability of betrayal, but rather, it is fixed (iii) *normatively* – viz., with reference to the (objective) normative concept of *de minimis* risk, viz., risks that can be *non-negligently* ignored by a truster.

The working idea that I will unpack, refine, and then put to work in order to handle our problem cases is the following:

> *De Minimis* Account (DMA): Trust essentially involves subjecting oneself to a risk of betrayal that is not merely *de minimis* within the relevant cooperative practice.

The starting point for unpacking DMA is that trusting as well as distrusting are both performative 'moves' within within the wider practice of cooperation – in a way that is roughly analogous to how belief and withholding are performative moves in practice of inquiry.[25]

'*De minimis*' is a normative term; a risk to the success of any performance (i.e., aimed attempt) is *de minimis* iff it can be *non-negligently* ignored in the course of making the relevant attempt.[26] *De minimis* risks are always *de minimis*, and thus have this normative standing, *relative to a practice*, where a given practice (i.e., a way of doing things) is held together by rules, either explicit or implicit.

What distinguishes the rules that *sustain* a given practice, as opposed to those rules that are merely incidental to it? A plausible general characterisation here, following John Turri,[27] is axiological: rules 'hold together' a practice whenever the *value* of following *those* rules explains why people engaged in that particular practice continue to follow them – viz., rules are practice-sustaining when they have 'reproduction value' within the practice.[28]

Accordingly, the initial idea that *de miminis* risks are practice-relative is

[25] See Kelp ("Theory of Inquiry," *Philosophy and Phenomenological Research*, 2020) for a recent defence of this kind of picture of inquiry.

[26] See Sandin ("Naturalness and de Minimis Risk," *Environmental Ethics* 27, no. 2 (2005): 191–200) and Peterson ("What Is a de Minimis Risk?" *Risk Management* 4, no. 2 (2002): 47–55) for discussion.

[27] "Sustaining Rules: A Model and Application," in *Knowledge First: Approaches in Epistemology and Mind*, ed. J. Adam Carter, Emma C. Gordon, and Benjamin W. Jarvis (Oxford: Oxford University Press, 2017), 259–277.

[28] See also Carter ("De Minimis Normativism: A New Theory of Full Aptness," *The Philosophical Quarterly*, 2020).

tantamount to the idea that *de minimis* risks – those that can be non-negligently ignored – will *have that normative status they have in connection with negligence always relative to a system of rules*, the rules that hold the practice together. And this, then, raises a question: in virtue of *what* would a given risk, e.g., to the success of *S*'s ϕ-ing, within a practice ψ, attain (when it does so attain) the normative status of being such that it could be *non-negligently* ignored by *S* with reference to the system of rules that constitutes ψ?

Here is a promising initial answer: we surely *can't* non-negligently ignore risks to the success of a performance within a practice *if there are rules with reproduction value within the practice, the following of which would easily mitigate against the risk.* (The archer, for example, can't non-negligently ignore whether the wind is blowing, even if the underwater diver can.)

But then – and this is the other side of the coin – we presumably *can* non-negligently ignore risks to a performance's success if the safety against that risk *can't* be easily increased through the truster's adherence to any rule whatsoever (e.g., "Monitor for this," "Check for that, etc.") that has reproduction value within the relevant practice. For example, the basketball player *can* plausibly non-negligently ignore risks of earthquakes prior to taking a shot, even though an earthquake would spoil that shot, given that monitoring for earthquakes lacks any reproduction value whatsoever in basketball (it is a rule the following of which would be a disvaluable distraction in the practice of basketball).[29] There are no rules with basketball reproduction value that a player *could* adhere to in order to easily safeguard against *that* risk. Thus, taking the non-obtaining of an earthquake scenario for granted is non-negligent during the making of performative moves within that particular practice, no matter how nearby the earthquake risk is modally: (after all, the *neglecting* of that possibility flouts no rules that are valuable to follow within the practice.)

Putting the key pieces together, we are now in a position to see how DMA works as a substantive answer to the Vulnerability Question we began with. DMA purports to answer that question by telling us in what sense trusting *essentially* involves subjecting oneself to risk of betrayal. And the answer to this question offered by DMA makes reference to the concept of risks that are *de minimis*, viz., risks that can be non-negligently ignored by a

[29] For a different explanation of *why* far-off risks such as the earthquake risk could be non-negligently ignored, see Sosa (*Epistemology* (Princeton: Princeton University Press, 2017), 191) and for more recent developments, Sosa ("Default Assumptions and Pure Thought," *Manuscript*, 2020).

truster. We have now got a working view of what such risks are and how to identify them: a risk of one's trusting being betrayed (alternatively: a risk to the success of one's trust) is *de minimis* and thus can be non-negligently ignored by a truster iff the safety of one's trust against that risk *can't be* increased through the truster's adherence to one or more rules that have reproduction value within the cooperative practice within which one is placing one's trust.

Let's now 'plug' this substantive characterisation of *de minimis* risk of betrayal back in to DMA in order to put the core idea of the view in full view. Since DMA maintains that trust essentially involves subjecting oneself to risk of betrayal that is *not merely de minimis* within the relevant cooperative practice, DMA tells us that trust essentially involves subjecting oneself to at least some risk or risks of betrayal that *aren't* merely *de minimis* – viz., that *aren't* such that one can't increase the safety against them by adhering to rules that have cooperative reproduction value within the relevant practice. That is the full way to spell out DMA – viz., that trusting involves rendering yourself vulnerable *beyond* mere *de minimis* risk of betrayal.

With the key components of the account now on the table, let's see what it can do, by checking whether it can – as advertised – fare better than the other accounts considered. Let's take, first, SUNRISE.

SUNRISE was not a case of *bona fide* trust. Our normative view DMA straightforwardly accommodates this. DMA says that trust essentially involves subjecting oneself to risk of betrayal that is *not merely de minimis* within the relevant cooperative practice. And the risk subjected to here is *de minimis* (i.e., it *can* be non-negligently ignored) because you can't increase the safety against *that* risk of betrayal by following any cooperation sustaining rule whatsoever. (Indeed, given the details of SUNRISE, this turns out to be *trivially* so; the likelihood of *that* risk event materialising remains the same (near zero) no matter *what* you do. Thus, by DNA, SUNRISE is not a case of genuine trust. So far, so good.

What about the BABYSITTER case? BABYSITTER *is* plausibly a case of trust, and a paradigmatic one, *despite* the very low objective probability of betrayal. If DMA is going to secure this result, then it had better be the case that you *could* at least in principle increase the (already robust) safety against risk to betrayal by adhering to rules with cooperative reproduction value. And indeed you can, and you can do so in relatively mundane ways: consider that such rules include vetting the babysitter *ex ante* (i.e., checking up on references for reliability), making babysitting itself easier (i.e., laying out emergency phone numbers, a list of medications, etc.): rules that encourage these have cooperative reproduction value. (Compare, by contrast: aiming to increase safety against the low risk present by *surveilling* the babysitter is non-cooperative; it is a form of monitoring we will discuss in the next section). Accordingly, then, the risk event that would consist in the babysitter failing to keep the child safe is not *de minimis* risk, even if it is very low due to the babysitter's impressive reliability and the straightforwardness of the task. Thus, DMA again gets the right result.

Let's turn now to PENCIL. This was also a case of trust, despite the very low albeit non-negligible *disvalue* of betrayal, and which generated a problem for an answer to the Vulnerability Question framed in terms of objective risk expectation value. For DMA to get the right result in this case, it had better bet that you *could* mitigate against the 'pencil theft' risk by the adherence to rules with cooperative reproduction value. And so you could. Writing you name on your pencil, for example, would violate no cooperation-sustaining rules; doing so facilitates rather than hinders cooperation between trustor and trustee. (More generally: the rule in play here would be to make items you loan out identifiable, which is a rule that has reproduction value for cooperation through loaning and borrowing). Accordingly, DMA is going to countenance PENCIL, rightly, as a case of trust, even though the disvalue of betrayal is exceedingly low (problematically so for the risk expectation value account to plausibly 'rule in' this case as a case of bona-fide trust).

The scoreboard of cases, then, is as follows:

Answer to Vulnerability Question	SUNRISE	BABYSIT.	PENCIL
> some threshold (T) of perceived risk	x	✓	✓
> some (T) of expected disvalue	✓	✓	x
> some (T) of (frequentist) prob. of betrayal	✓	x	✓
> (normative) *de minimis* risk	✓	✓	✓

II.d. Objections and replies

So far, it is looking liker DMA outperforms the competition as an answer to the Vulnerability Question, at least in so far as the view gets on the wrong side of none of the three cases that posed a problem for at least one of the other views considered. This is a promising mark in favour of

DMA. Let's now see how the proposal holds up against some anticipated objections.

Objection 1 Even if we grant that the DMA gets the SUNRISE case right, there are nonetheless 'Frankfurt-style' cases (also with effectively zero risk of betrayal) that pose a problem to *any* view that takes some (non-zero) objective risk of betrayal to be necessary for trust.

Notice that it is a feature of SUNRISE that, given the effectively zero objective chance of betrayal, which would have involved moving the earth's orbit, it was completely *out of the control* of the trustee whether they betray or not, such that betrayal (or not) isn't something that could be attributed to them. Even so, it seems like we can imagine cases where the following *both* hold: (i) there is zero objective chance of betrayal; but (ii) where it is *not* out one's control whether they betray or not such that we could attribute *at least trust fulfilment* to the trustee, thus satisfying the weak attribution principle. For example, consider FRANKFURT-BABYSITTER:

> FRANKFURT-BABYSITTER: Suppose this case is just like BABYSITTER, except that it is a Frankfurt case[30], in that *if* the babysitter were to do something that would in any way imperil the baby, a benevolent demon would rush in and course-correct, preventing any danger to befall the baby. Because the babysitter (through her own goodwill and reliability) does everything right, the benevolent demon never has to intervene.

If we are going to retain the idea that trust essentially involves 'some objective risk' of betrayal – a conclusion from the critique of the perceived risk account – it looks like we're going to get the wrong result, i.e., that this isn't a case of trust. But, as the worry goes, it *is* trust despite there being no objective risk whatsoever that the babysitter will *not* come through. So long as she behaves in such a way that the Frankfurtian demon needn't intervene, all is good.

Reply It is important to distinguish (i) risks to *successful reliance* and (ii) risks to *successful trust.* If you rely on someone to ϕ, your reliance is successful iff they ϕ, no matter how.

Trust asymmetrically entails reliance. When you trust someone to ϕ, you

[30] See, e.g., Frankfurt ("Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66, no. 23 (1969): 829–39).

trust them to ϕ *as entrusted*, where 'as entrusted' might include such things as: with goodwill toward the trustor,[31] by encapsulating the interests of the trustor,[32] by believing they have a commitment to the truster to ϕ,[33] etc.

For my purposes, I am happy to remain neutral on which of these ways of unpacking 'as enstrusted' best distinguishes trust from mere reliance. What is relevant at present is just that the success conditions for reliance and trust differ, in that trusting someone to ϕ is successful iff they ϕ *as entrusted* (however this is to be spelled out), and not *merely* iff they ϕ.

This difference in success conditions is important in defusing the above objection. This is because *there can be risks to successful trust that are not also risks to successful reliance.* And indeed, that is exactly what is going on in FRANKFURT-BABYSITTER. It is true that there is zero objective risk to successful reliance; the benevolent demon waiting in the wings is seeing to that. But it is not *thereby* also true that there is zero objective risk to successful *trust.* The trustee's taking care of things as entrusted – however we fill this out – is plausibly going to require some exercise of autonomous agency, some *way* of taking care of things, attributable to the trustee – a point that lines up with the observation that reactive attitudes like gratitude are appropriate to fulfilled trust as well as to betrayal. Actions and mental states caused by the demon's intervention are not autonomous[34]; when compelled to act by the demon, the agent is not free to govern herself one way or the other, with respect to what she has been entrusted to do. She *cannot* fulfil trust (even if she can play a causal role in bringing about what she was relied on to do) or betray it.

Thus, there *is* some non-zero objective risk of betrayal (i.e., a risk to successful *trust*) in FRANKFURT-BABYSITTER, even though there is no objective risk to successful reliance. And importantly, the objective risk to betrayal is (as is pertinent to DMA) not *merely de minimis*: This is because, just as there are pro-cooperative rules you could adhere to to increase safety against risk of betrayal in the original (non-Frankfurt) BABYSITTER case, so likewise, the same applies here – viz., such rules include,

[31] E.g., Baier "Trust and Antitrust"; Jones "Trust as an Affective Attitude," *Ethics* 107, no. 1 (1996): 4–25.

[32] E.g., Hardin *Trust and Trustworthiness* (Russell Sage Foundation, 2002).

[33] E.g., Hawley "Trust, Distrust and Commitment," *Noûs* 48, no. 1 (2014): 1–20.

[34] There are different ways to explain why. For two prominent options, see, e.g., Mele (*Autonomous Agents: From Self-Control to Autonomy* (Oxford University Press on Demand, 2001)) and Fischer and Ravizza (*Responsibility and Control: A Theory of Moral Responsibility* (Cambridge university press, 2000)).

e.g., proper vetting, and (post-vetting) facilitating cooperative attitudes of the trustee through, being cooperative as a truster – e.g., by making duties clear. DMA therefore is able to handle not only BABYSITTER but also FRANKFURT-BABYSITTER.

Objection 2 Let's consider now a further objection to DMA, one that serves well as an entry point into the discussion in the next section on the relationship between trusting and monitoring.

Consider that on the proposal advanced, trust essentially involves subjecting yourself *beyond* mere *de miminimis* risk of betrayal – that is, it essentially involves subjecting yourself to at least some risks of betrayal that can't be non-negligently ignored – viz., such that the safety against them *couldn't* easily be increased through adherence to rules with cooperative reproduction value. But a corollary of this idea is that that *all* cases of trust are ones where you could at least potentially increase safety against betrayal by following rules with cooperative reproductive value.

But – and here is the worry – isn't *this* commitment of the view somehow in tension with the platitudinous idea that trusting is incompatible with *monitoring*? After all, there will often be no more effective way to increase safety against betrayal than to blatantly monitor the trustee's every move. Rather than to, e.g., mitigate against betrayal by carefully vetting the babysitter's references and then leaving helpful reminder notes, why not simply watch the entire time with surveillance cameras, or – better yet – hire a full surveillance team to oversee the babysitter's every move?

In short, the objection to the proposal can be put like this: the answer given to the Vulnerability Question – viz., that trust essentially involves subjecting yourself beyond mere *de miminimis* risk of betrayal – rests on the underlying idea that when one trusts, there *are* certain things that one can do to increase safety against betrayal. But, increasing safety against betrayal is (at least in cases of monitoring, which is one very obvious way to increase safety against betrayal) incompatible with genuinely trusting. Thus, it seems that the answer given to the Vulnerability Question cannot be satisfactory: it relies on a claim that is itself in tension with the datum that monitoring kills trust.

Reply This is a straightforward objection, and it has an equally straightforward answer. *Monitoring*, even though it increases – perhaps better than anything else! – safety against betrayal, is fundamentally *non-cooperative.* For one thing, that norms of cooperation generally prohibit monitoring or surveilling a trustee is supported by our practices of sanctioning; we tend to sanction those who purport to trust and then monitor.[35] Additionally, monitoring contributes to the erosion of conditions for cooperation; this is due to the social function of monitoring as *signalling* a lack of confidence in a pre-established commitment.[36]

Importantly, the view advanced here does not maintain that by trusting you subject yourself to risks of betrayal such that you could (while continuing to trust) *in any way* increase the safety against their obtaining. *That* would indeed be an unacceptable result. It implies rather that trusting essentially involves subjecting yourself to risks of betrayal that are not merely *de minimis*, which just means that by trusting you subject yourself to at least some risks of betrayal such that you could (in principle, and regardless of whether you do) increase the safety against their obtaining without violating any cooperation-sustaining rules.

And indeed we increase safety against betrayal without violating any such rules like this *all the time* (and *without monitoring*): by deliberating about whom to trust, assessing their reliability, assessing facts pertinent to the likelihood of betrayal, including the extent, present in a given trust context, of the (a) *gains to the trustee* that would come from betrayal; (b) the *effort*; and (c) the *aptitude* required by the trustee to *avoid* betrayal. Through a competent assessment of these factors one can cooperatively increase safety against risk of betrayal. (Likewise, one can cooperatively increase safety against risk betrayal by facilitating the ease by which the trustee can take care of what is entrusted, e.g., by leaving a map, leaving detailed instructions, etc. None of these things involves trust-incompatible monitoring.)

In sum, then, the idea that trusting essentially involves subjecting yourself beyond mere *de minimis* risk of betrayal does not stand in tension with the platitude that trusting is incompatible with monitoring the trustee.

III. TRUST AND MONITORING

The final objection in the previous section prompts a further question - in what sense, then, is monitoring for risks of betrayal incompatible with trusting? This is the Monitoring Question. The Monitoring Question

[35] See, e.g., Kramer ("Trust and Distrust in Organizations: Emerging Perspectives, Enduring Questions," *Annual Review of Psychology* 50, no. 1 (1999): 569–98).

[36] For some studies reporting these effects in cases where computers are used to monitor employees, see Ariss ("Computer Monitoring: Benefits and Pitfalls Facing Management," *Information & Management* 39, no. 7 (July 1, 2002): 553–58).

takes at face value that monitoring *is* incompatible with trusting. It invites us to explain *how so*.

Just as a good answer to the Vulnerability Question required some sense of what the threshold is beyond which by trusting one subjects oneself to risk, a good answer to the Monitoring Question requires some sense of what the threshold is beyond which by monitoring one is no longer trusting.

Here is the answer to the Monitoring Question I will now defend:

> (MON): One's monitoring is incompatible with trusting to the extent that, through monitoring, one intentionally aims (through the taking of some means) at *invulnerability* to risks of betrayal that, by trusting, one essentially subjects oneself to.

Since by trusting one essentially subjects oneself beyond mere *de minimis* risk of betrayal, MON implies that monitoring is incompatible with trusting to the extent that it involves taking means by which one aims to render oneself *invulnerable* to all but *de minimis* risks of betrayal.

Two initial clarifications here are needed. First, the proposal does not say that one *actually* has to render herself invulnerable to all but *de minimis* risks of betrayal. This is important, because the monitoring needn't actually succeed in that aim to be incompatible with trust. Consider, for example, the following case:

> CRYSTAL BALL: *A* hires *B* to babysit *A*'s child. Highly superstitious, *A* believes, falsely, that *A* has a working crystal ball. After dropping *A*'s children off with B, A hurries home to the crystal ball in an attempt to surveil *B*'s every move. The crystal ball shows ambiguous, smoky images, which *A* mistakenly thinks provide information about *B*'s movements. *B* watches and attempts to interpret these movements, much as a less superstitious person might peer into the grainy images on a nannycam with poor resolution.

Intuitively, *A* is no longer trusting *B* when using the crystal ball – anymore than one surveilling via a poor-resolution nannycam would be doing so – and *even though A* is not succeeding in making herself invulnerable to any risk of betrayal whatsoever. An account of the incompatiblity of monitoring with trusting that required actually succeeding in eliminating, even to some degree, such vulnerability would fail to get the right result in

CRYSTAL BALL.

A second clarification: why 'the taking of means by which one aims?' Why not *simply* 'aims?' The reason is that monitoring – as opposed to something less, i.e., merely intending but failing to monitor – requires an attempt to *attain* an aim (i.e., vulnerability elimination) *in some way*, viz., through some means by which one *through taking those means* (in this case, via the means of surveilling the trustee) monitors; in this respect, monitoring that is incompatible with trusting is not 'idle aiming' (i.e., mere intending to monitor) any more than trusting is idle aiming (mere intending to trust).

These clarifications made, we can see now that MON is able to secure the following pleasing result: it can explain why surveilling the babysitter with a nannycam (or, for that matter, attempting to do so via a crystal ball) is incompatible with trusting but vetting the babysitter for reliability (prior to hiring) and leaving notes and reminders after is not, even though the latter kinds of things also minimise risk of betrayal. The explanation given is that, in the former examples, one intentionally aims (through the taking of some means) at invulnerability to risks of betrayal that, by trusting, one essentially subjects oneself to (i.e., to risks that aren't merely *de minimis* in the relevant contexts), whereas this is not so in the latter cases.

The fact that MON is able to generate different verdicts in the former and latter kinds of cases constitutes a key advantage over a more standard line of thought about trusting and monitoring in the literature[37] according to which trusting essentially involves simply refraining 'from taking precautions against an interaction partner.'[38] Such a proposal frames the relationship between trust and monitoring in a problematically coarse-grained way; it would get the former cases right, but not the latter unless it could provide us (as MON does) a principled reason for a difference in treatment.

An additional advantage of MON is that the very idea that 'aiming' at eliminating vulnerability is something that is incompatible with trusting fits snugly with a much more basic idea about trusting *qua* performance, or aimed attempt. Within the theory of performance normativity, performance types may be distinguished from each other by the constitutive aims internal to those performance types. In slogan form: change the

[37] E.g., Elster *Explaining Social Behavior* (Cambridge University Press, 2015).

[38] 344.

aim, and you've changed the performance type.[39] Take for example the performance of 'making a guess' versus 'making a judgement'; these are different truth-directed performances; but why? A typical answer[40] adverts to a difference in the level of risk one aims at taking on as a price for a chance at truth in each case. *What it is* to make a guess is to aim at truth via affirmation in a way that tolerates at least an unusually high level of risk; were one to aim at truth via affirming *without* aiming at tolerating whatever level of risk is distinctive of guessing, then one is longer guessing. The idea is that the same goes for trusting in so far as by trusting we aim at something in a way that essentially renders us vulnerable. Monitoring a trustee (by intentionally aiming to immunise oneself from such vulnerability) alters this aim distinctive of trust, changing the performance in a way that is broadly analogous to how (through the process of collecting more evidence) one is no longer guessing, but believing.

IV. CONCLUDING REMARKS

The principal objective here has been to defend new answers to the Vulnerability Question and the Monitoring Question, answers shown to fare better than the competition. But in doing so, I've also tried to uncover an important but unnoticed way in which these questions are connected to each other, with the Vulnerability Question the more fundamental of the two. On the view defended, which views both questions through the lens of performance norms, monitoring a trustee is incompatible with trusting to the extent that, through monitoring, one intentionally aims at invulnerability to risks of betrayal that, by trusting, one essentially subjects oneself to (§III). But which risks are these? It is at this point that our answer to the Vulnerability Question kicks in: trusting essentially involves rendering oneself vulnerable to betrayal in the sense that it essentially involves subjecting oneself to risk of betrayal that is not merely *de minimis* within the relevant cooperative practice (§II.a). And – putting these ideas together – the fuller answer to the Monitoring Question, framed in terms of our answer to the Vulnerability Question, is that monitoring is incompatible with trusting insofar as one intentionally aims at invulnerability to not merely *de mimimis* risks of betrayal, viz., to not *merely* those risks to which, by trusting, one essentially renders herself vulnerable.

[39] For discussion of how performances are individuated by their aims, see Sosa ("How Competence Matters in Epistemology," *Philosophical Perspectives* 24, no. 1 (2010): 465–75), and the essays in (ed.) Vargas (*Performance Epistemology: Foundations and Applications* (Oxford University Press, 2016)).

[40] See, e.g., Sosa (*Judgment & Agency* (Oxford University Press UK, 2015), Ch. 3).