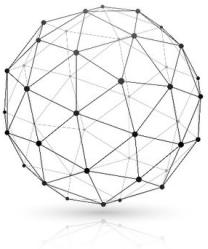
BlockchainGov



In Blockchain We Trust(Less):
The Future of Distributed Governance

READING GROUP

The Blockchain Governance Reading Group will meet every two weeks to discuss a book chapter or an article related to some of the key questions that will be addressed as part of the ERC Blockchain Gov project.

Every participant to the reading group is expected to have done the readings. The session will start with a 10 minutes presentation and commentary from a discussant, followed by an open discussion among the participants who will express their views and opinions with regard to the selected readings. When possible, the author(s) will be invited to join the call in order to get their immediate reactions.

The reading group will be held every two weeks, on Thursday at 6pm CET (starting January 14th). Each reading group will last for a period of two to three months. The initial readings have already been identified, the others will be decided on an on-going basis by the reading group participants.

The following themes will be addressed through the coming year:

- *Blockchain, Trust & Confidence* (Jan-March 2021)
- *Blockchain, Coercion & Legitimacy* (April-June 2021)
- *Polycentric governance* (July-Sept 2021)
- *Blockchain & the Rule of Law* (Oct-Dec 2021)

Session I: **Blockchain, Trust & Confidence**

The objective of this first session will be to explore ways in which blockchain technology could contribute to bringing more trust and confidence in existing social and institutional arrangement. Readings will cover the notions of trust and confidence, and will investigate the extent to which blockchain technology can either strengthen or weaken these concepts. The goal is, ultimately, to have a better understanding of how to design blockchain-based systems that promote trust and confidence.

Agenda: **Thursdays. bi-weekly at 6pm** CET

- **Jan 14** —Luhmann (2000): Trust & Confidence
- **Jan 28** —Gambetta (2000): Can we trust trust?
- **Feb 11** —Hardin (2002): Trust & Trustworthiness. Chapter 1: Trust
- **March 2** —Nissenbaum (2004): Trust as Security (Guest speaker)
- **March 11** —Balazs (2020): Mediated Trust (Guest speaker)
- **March 24** — Schneier, B. (2012). Liars & outliers, Chapter 6 (Guest Speaker)
- **April 8** — Pettit (1995): Cunning of Trust; (2004): Trust, Reliance & Internet
- **April 22** — Judith Donath (current draft of her book on Trust)
- **May 6th** — Coeckelbergh (2012). Can We Trust Robots? (Guest Speaker)
- **May 20th** —Patrick Sumpf (2019) System Trust: Researching the Architecture of Trust in Systems —Chapter 2 (Guest Speaker)
- **June 10th** —Simon (2010). The entanglement of trust & knowledge on the Web (Guest speaker - discussant: Michael Heidt)
- **June 17th —** Nguyen - "Trust as an unquestioning attitude" (Guest Speaker)
- **July 1st —** Sunshine (2007). Public Confidence in Policing
- **July 15th—** Fran Tonkiss, Andrew Passey (1999) - Trust, Confidence and Voluntary Organisations: Between Values and Institutions

Ideas for the reading group on *legitimacy: starting in August*

- **August 19th—** Amanda Greene - Consent and Political Legitimacy
- Marcia Grimes - Organizing consent: The role of procedural fairness in political trust and compliance
- Arthur Applbaum - Legitimacy without the Duty to Obey
- Rainer Forst - Normativity and Power, Chapter 8
- Allen Buchanen - Political Legitimacy and Democracy
- Fabienne Peter - Democratic Legitimacy and Proceduralist Social Epistemology
- Ian Hurd - Legitimacy and Authority in International Politics

- Sofia Näsström (2007). - The Legitimacy of the People

ZOOM LINK: https://sciencespo.zoom.us/j/8789855081

Invited Participants:

- **Blockchain Gov team***: Primavera De Filippi, Philémon Poux, Nicholas Gouverneur, Wessel Reijers, Morshed Mannan, Simona Ramos, Paula Berman, Jack Henderson*

- **Scholars***: Andrea Leiter, Beatriz Botero, Charles Nesson, Juan Ortiz, Nick Couldry, Quinn Dupont, Vasilis Kostakis, Georgy Ishmaev, Silvia Semenzin, Giovanni Sartor, Gianluca Miscione, Geert Lovink, Mattis Jacobs, Lana Swartz, Bruce Schneier, Balazs Bodo, Liav Orgad, Elettra Bietti, Sandra Braman, Inte Gloerich, Judith Donath, Vera Shikhelman, Jessy Kate Schingler, Mireille Hildebrandt, Rafael Zioloswki, Divya Siddarth, Bruno Deffains, Liudmila Zavolokina, John Danaher, Nathan Schneider, Brett Frischmann, Chris Wray, Matt Prewitt,* Eric Alston*, Yochai Benkler, Victoria Lemieux, Larry Backer, Ori Freiman, Michael Heidt, Christopher Ba Thi Nguyen, Rachel O'Dwyer, Gili Vidan, Samer Hassan, Kevin Werbach, Sheila Jasanoff, Lawrence Lessig, Saskia Sassen, Xin Dai, Jason Potts, Davidson Sinclair, Darcy Allen, Chris Berg, Shawn Bayern*

Session I. Jan 14 —Luhmann (2000): Trust & Confidence

Attendants: Primavera De Filippi, Wessel Reijers, Morshed Mannan, Nick Couldry, Judith Donath, Paula Berman, Balazs Bodo, Vasilis Kostakis, Inte Gloerich, Nicholas Saul, Andrea Leiter, Quinn Dupont, Silvia Semenzin, Georgy Ishmaev, Simona Ramos, Philémon Poux, Mattis Jacobs, Chris Wray.

Key concepts

- Trust and confidence refer to expectations which may lapse into disappointments. [very different disappointments: disappointment in probability changes nothing but further probability; disappointment in belief is a whole different thing.]

- Trust and confidence are placed in a familiar world by symbolic representation, and therefore remain sensitive to symbolic events. [treating them like identical twins does no service to their distinction]

- Modern times and liberalism introduce a shift from cosmology to technology, from religion to law and politics, from "fortuna" to "risk". [reshaping human identity]
- The printing press has also changed our modes of coping with the unfamiliar, by making it possible to pursue knowledge and risk assessment. [words enable us to think, record, compute]
- The distinction between confidence and trust depends on perception and attribution. [perception associates with intake of data; attribution to belief]
- In the case of confidence you will react to disappointment by external attribution.
- In the case of trust you will have to consider an internal attribution and eventually regret your trusting choice.
- Political and economic liberalism attempts to shift expectations from confidence to trust.
- The increasing complexity of social systems requires increased confidence in order to economise resources.

Discussion (Wessel Reijers)

- Paper was triggering both interest and confusion.
- Unity between familiar and unfamiliar—is it too hegelian?
- Trust & Confidence defined as expectations that might lapse into disappointment—but expectations not well characterized.
- Risk taking as sustaining the process of production and consumption—doesn't this characterization of trust and confidence relies too much on capitalist liberal paradigm?
- Distinction between religion / law & politics. But does it hold? What about theocracy?
- Isn't trust reduced to calculation?
- Can't Luhmann's definition of confidence be applied to pre-modern times?
- If dangers happen on the macro-level, while risks are taken on the micro-level–what does it mean to have confidence on experts in large, macro questions like climate change?

Comments

- *Chris Wray:* Repeated references to systems theory, but claims seem imprecise.
- *Nick Couldry:* Paper is a puzzling mixture of different things. Micro differences between concepts. The functionalism that the author hints at comes in a big way: trust is about trust in the system. But Luhmann never explains how a system like that is built, what is the system based on? It is not clear what the author means by "the system" or "functional

systems"? (Nick is) Allergic to systems theory, but agrees with the author's point that trust is very important in social order, and distinct from confidence. However, there's a *human hole* in his paper. Other forms of human trust and agency are absent. Similar to Primavera's point that humans are key to proper blockchain functioning, but also absent from "blockchain governance" analysis.

- *Primavera De Filippi:* The author is not trying to look at how systems are built, but what is the relationship between the individual and the system. Author points out that these relationships are always about expectations. You are never sure, but in "trust" you know that you are making a specific choice, and that's where the risk comes from. Whereas in confidence, you don't choose, you just are confident, while acknowledging that there are dangers. You can trust the medical license that the scientific community gave to a doctor, so you see the doctor.

- *Quinn Dupont:* Illuminating passage, page 99 part III. Author is not answering where systems come from but pointing out that *trust* and *confidence* are prerequisites in socially complex systems. Paper has an interesting metaphor about the origins of money, which are important for blockchain research.

- *Andrea Leiter:* Two reactions to the text: i) The historical contingency that he is pointing to was very interesting. While Luhmann is very protective about his categories, he drove the argument that they work through different historical contingencies; ii) By tying trust with an element of attribution, or action, and risk, the author is dealing with the question of control. The angle that comes through is: how can I control my environment, when giving up cosmology? *Luhmann acknowledges unknowability, contains it and makes it tameable.* To her, in our current times where there are many questions over human-mastership of everything, it's more interesting to look at how to make the environment livable, but not necessarily controllable.

- *Primavera De Filippi:* Confidence & trust are highly subjective: up to the individual to decide. The more I understand something, the more I can trust it, or not. The paper is essentially about how we relate to the external world. Before we had to trust religion, but now with science, confidence emerges. But then again, trust is necessary because we trust the experts behind science. This means that the system of confidence, unless it's driven by personal experience, requires an underlying element of trust.

- *Simona Ramos:* The point is whether or not disappointment depends on previous behaviour. It's essentially about our ability to have a say, or not, within systems.

- *Morshed:* Pre-modern societies existed in a state of trust, and modernity is about creating more alternatives. Interesting to look at theocratic states, where to this day certain types of societies, that present fewer alternatives to trust, require enormous amounts of confidence (not in a positive sense). Genesis block as an example of both trust and confidence eroded. Blockchain as a way to rebuild confidence but through a different institution.

- *Primavera De Filippi:* When there is too much complexity, trust becomes necessary. I could do it myself but I should trust others to not deal with too much complexity.

- *Philémon Poux:* familiarity leads to confidence, is necessary at the beginning, but then disappears with more complexity. Important to think about familiarity. Distinction between what lack of trust leads to (smaller bubbles) or lack of confidence leads to. We are observing today more than ever, these smaller bubbles.
- Judith Donath: Understands trust as emerging out of relationships. A lot of what we do are ways of establishing trust with each other. Finds confusing that confidence isn't well defined in the paper. She thinks about "confidence" in terms of particular circumstances, it's situation-dependent. Blockchains can increase confidence or reduce risk and thus reduce or eliminate the NEED for trust.
- *Vasilis Kostakis:* Trust is two-fold: it may be a means to an end (end=confidence) or the end itself (trust for the sake of trust; because we need to feel human)
- *Balazs Bodo:* Paper lacks integration of concepts of familiarity, trust and confidence. Where does the behavior of the DAO members fit?
- *Primavera De Filippi:* There was the aspect of risk of a traditional investment, but most people were confident that the smart contract would behave as planned.
- *Balazs Bodo:* In studying trust and confidence it's important / helpful to look into situations where you can clearly identify agencies and risks.
- *Simona Ramos:* Range of participation can be informative in analysing trust or confidence in the blockchain setting: miners may be in a situation of trust whereas end users act on confidence.
- *Charles Nesson:* The distinction and relationship Luhmann seeks to make between trust and confidence mirrors the distinction and relationship between belief and probability as used in law. Is the jury's verdict of criminal guilt an assertion to the community at large of belief or probability? Contrast the jury's verdict of liability in a civil trial. Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts,* Harv.L.Rev. *(1985)*

Session II. Jan 28 —Gambetta (2000): Can we trust trust?

Attendants: Primavera De Filippi, Wessel Reijers, Morshed Mannan, Judith Donath, Paula Berman, Balazs Bodo, Inte Gloerich, Nicholas Saul, Andrea Leiter, Quinn Dupont, Georgy Ishmaev, Chris Wray, Gianluca Miscione, Divya Siddarth, Beatriz Botero, Charles Nesson, Juan Ortiz, Jessy Kate Schingler.

Key concepts:

On rational cooperation:

- At surface level, *rational motives or interest* are fundamental for cooperation, yet at closer inspection *beliefs* (i.e. trust) play a key role:
    - first degree trust: our beliefs about others, w.r.t their willingness to cooperate;
    - second degree trust: our ability to establish our own trustworthiness, so others believe we will cooperate with them, and will therefore cooperate with us.
- Thus, rationally motivated cooperation may not emerge, not due to absence of rational motives but more simply because *not enough people trust others to act by those motives.*

Definition of trust:

- Trust is defined as a threshold point, located on a *probabilistic distribution of subjective expectations*, ranging from complete distrust (0) and complete trust (1), and which is centred around a mid-point (0.50) of uncertainty.

Conditions under which trust becomes relevant for cooperation:

- Trust entails *ignorance:* it is a way of coping with information asymmetry:
    - With unlimited computational ability to map all possibilities, trust would be unnecessary
- Trust entails *our freedom:* in absence of freedom to choose there can be no trust, only hope.
    - Thus, for Gambetta, "confidence" in Luhmann's sense is akin to hope, or blind trust.
- Trust entails *the freedom of others*: trust is relevant only if one party can disappoint the other
    - it becomes more salient, the larger the feasible set of alternatives open to others.
    - we can increase trust by means of commitment, coercion, contracts, promises, etc.
    - *yet, coercion is not an exhaustive "functional equivalent" of trust:*
        - it is more costly (expenditures on surveillance, information gathering and enforcement)
        - “imposed coercion” might reduce trust in the party exercising such coercion, thereby increasing the probability of defection by the coerced.

    - pre-commitments to impose restraints on our action and reduce the set of alternatives

- Trust can be *artificially constructed*, if the alternatives are worse or interests are very large
    - if the pressure to act is great even when the trust threshold is lower than 0.50 – when we verge on distrust; by means of wishful thinking and the reduction of cognitive dissonance, a deceptive rearrangement of beliefs can be generated.
- The demands for trust on cooperation is circumscribed by the *constraints, costs and benefits* presented by each specific situation.

  ⇒ Trust can be a result of, rather than a precondition for cooperation [more confidence = more trust]

  How to promote cooperation, without relying on trust:

- *Economizing on trust:* standard practice of manipulating constraints, costs and benefits.
    - cooperation via coercion + arrangements that encourage cooperation for self-interest
- *Signalling shared interest:* ppl are unlikely to cheat if that would go against their interests
    - *Signals* emitted by the multiple agents, pointing to shared, mutual interests.
    - Mutual interests make defection costly enough to be deterred [e.g. Mining]
- *Evolutionary perspective:* optimal cooperation can evolve independently of trust, as in the animal world and in certain human settings (e.g. war zones).
    - Problem: in nature, such a process happens randomly, thus cannot be relied upon. Further, without an initial *predisposition to trust,* it is not likely to evolve.
- *Unilateral blind trust*: game theory strategy to generate a productive *tit for tat* dynamic.
    - Problem: seemingly trusting signal could be interpreted as a trap. Thus, tit for tat can be an equilibrium only if both players *believe* (trust) the other will abide by it, otherwise other equilibria are just as possible and self-confirming.

  ⇒ it might be rational to act "**as if**" we trusted the other party, in order to establish the initial grounds for cooperation and therefore increase the likelihood for real trust to be established among the players
  = trust as a choice, rather than a fortunate by-product of evolution

  Definition of trusting trust and distrusting distrust:

To choose deliberately a testing value of p which is both high enough for us to engage in tentative action, and small enough to set the risk and scale of possible disappointment acceptably low.

Trusting Trust as a rational strategy:

- There is no evidence for "trustworthiness" (only lack of evidence for breach of trust)
- Trust begins with keeping oneself *open to evidence,* acting as if one trusted, at least until more stable beliefs can be established on the basis of further information. If we do not act *as if we trust*, we shall never find out if our suspicions were justified.
- When we behave *as if* we trusted the other party, we *enable cooperation* through *conditional trust*, which is based on the belief that the other party is not a sucker, but also on the belief that he will be well disposed towards us if we make the right move.
- *Reputation for trustworthiness* is a very valuable commodity:
    - Interests might create incentives to act honestly, but reputation and commitment are the way in which others are assured of the effectiveness of that pressure
- Trust is not a resource that is depleted through use; on the contrary, *the more there is the more there is likely to be.*
    - When trust is not unconditionally bestowed it may generate a greater sense of *responsibility* at the receiving end. The concession of trust, that is, can generate the very behaviour which might logically seem to be its precondition.
    - Trust is depleted if it is not sufficiently used
    - If behaviour spreads through learning and imitation, then sustained trust can lead to trust, and sustained distrust can only lead to further distrust.

Distrusting distrust as a rational strategy

- Similarly, "distrust may become the source of its own evidence" : distrust encourages us to look for evidence to disprove trust, hence creating a self-fulfilling prophecy where we become increasingly more distrustful
- Moreover, acting in terms of distrust may generate more distrust from others
- The alternatives to banking on trust can be so *drastic, painful, and possibly immoral* that they can never be lightly entertained.

- Being wrong is an *inevitable* part of the wager, of the learning process strung between success and disappointment, where *only if we are prepared to endure the latter can we hope to enjoy the former.*

- Asking too little of trust is just as ill advised as asking too much.

Discussion (Nicholas Saul)

- A very stimulating piece that brought me back a few years to some of my initial research on game theory, collective intention and the notion of subjectivity.
- Methodological individualism is at the core of the debate: can we trust trust purely from the point of view of a subjective rational motive or interest *and/or* belief?
- Is trust really *only* a **probabilistic distribution of subjective expectations**, and if so how do we categorize collective intentions, beliefs etc. What do we do of the plural subject? How does the plural subject create trust?
- Fundamentally my real question comes to the idea of whether or not trusting trust and distrusting distrust can be limited to a rational strategy or whether we need to encompass other forms of intentionalities? (I ask this as an open question in relation to what we consider as co-governance of the commons)
- Researchers like Margaret Gilbert consider this limited : individualists think we mustn't go beyond the individual category of desires, beliefs, objectives and engagements to « describe the context in which human beings deliberate and act ». Examples of such contemporary individualists : H.L.A. Hart, David K. Lewis, Michael Bratman
    - What can the theory of plural subjects and joint engagements bring to this conversation? The links and attachments that unify forms of life, and allow for a governance of the commons, are they not essential social phenomena to take in account beyond individual belief when talking of trust?
- All the examples used seem to describe a game theory temporality of confrontation and gain situation (i.e. war, arms race, individual confrontation (prison dilemma), etc.)
- Argument on animals, an example that cooperation is a sympoietic / intersubjective relation?
- Question of the varieties of learning and intentionalities => how do they depend on the varying forms of cooperative practices? (example of markets and states => welfare born from this? Notion of aristocracies *confianza en confienza*). Does trust follow rather than precede cooperation?
- The prisoner's dilemma issue according to Axelrod's *Tit for Tat* strategy (generalizing the evolutionary approach?): thinking only in the axis of confrontation & gain: not the way ordinary life or commons-based governance can be obtained (i.e. through the rules & norms (technical or ordinary) under which the users of a community co-govern the ressources.)
- Intentionality must be invoked, to what extent can it be invoked, but especially: what kind of intentionality? Diego Gambetta touches upon this p.14
- If behaviour spreads through learning and imitation then is it not the types of intentionality that we learn and imitate that need to be changed in order to trust trust?

Meeting Agenda

- *(name): impressions / topics for debate*

- *(Juan Ortiz Freuler):*
    - *In referencing the work of Desgupt and Lorenz, you state "If we were blessed with an unlimited computational ability to map out all possible contingencies in enforceable contracts, trust would not be a problem"*
        - *Though this seems sound, is it not also true that trust creates value beyond cooperation?*
        - *What is your take on replacing trust with trustlessness? Would that not create negative side-effects on the social bonds?*
    - *You state "Contracts and promises represent weaker forms of pre-commitment, which do not altogether rule out certain actions, but simply make them more costly. Contract shifts the focus of trust on to the efficacy of sanctions"*
        - *Relying on the definition of contracts, given "smart contracts" are self-executing, and leave no space for deviation, should we not choose another term?*
    - *I agree that dysfunctional societies of the south ["underdeveloped"] typically lack cooperation, yet perhaps this lack of cooperation could be traced back to colonial structures of extraction that relied on fueling internal divisions to cement their control, and minimize chances of successful uprisings?*

- (Primavera De Filippi): reflections on how the key points of Gambetta (highlighted in purple below) relates to blockchain technology:

    - *if unlimited computational ability to map all possibilities, trust would be unnecessary*
      in that case, trust would be replaced by confidence (as per Luhmann definition)
      what about the performative functions of trust? (e.g. create stronger social bonds)

    - Trust entails **the freedom of others**: relevant only if a party can disappoint the other
        - *trust is more salient, the larger the feasible set of alternatives open to others.*
        - *can increase trust by means of commitment, coercion, contracts, promises,*

      c.f. blockchain as a "trustless system" where everything has been pre-determined

    - *pre-commitments to impose restraints on our action and reduce the set of alternatives*
      "smart contracts" can be regarded as a means for parties to impose self-constraints, with a view to increase the confidence in the way in which the transactions will run

    - Trust can be a result of, rather than a precondition for cooperation
      more confidence = more room for building trust relationships on top ?

    - cooperation via coercion + arrangements that encourage cooperation for self-interest
        - ppl are unlikely to cheat if that would go against their interests
        - Mutual interests make defection c
        -
        - ostly enough to be deterred

      cf. PoW / PoS mining as game theoretical design to encourage cooperation

    - Reputation **for trustworthiness** is a very valuable commodity:
        - Interests might create incentives to act honestly, but reputation and commitment enable others to be assured of the effectiveness of that pressure

      e.g. reputation and token staking, to increase trustworthiness in a blockchain system

    - There is no evidence for "trustworthiness" (only lack of evidence for breach of trust)

      hard in pseudonymous systems, in which people can split identities for misbehaviour

- (Divya Siddarth): How do these relationships change in situations of significant power imbalances between parties? Many of the concrete examples are to an extent 'between equals' - soliders on different sides, networks of moneylenders. Can we expand the discussion of tension betweeen intensity of interest in action and the value of p - for example, how does the concept of 'trusting trust and distrusting distrust' play out when one party has significant distrust but no leverage, and the other party has no incentive to cooperate - ie in precarious labor relationships between employers and gig workers, or between debtor countries and the IMF?

Transcript:

- (Primavera de Filippi) Interested in the question of whether trust is a precondition or consequence of cooperation.

- (Primavera de Filippi) When relying on trust, cooperation entails creating conditions where not only I need to trust you, but I need to trust that you trust me to trust you.

- (Primavera de Filippi) Interesting correspondence between what Gambetta writes in terms of pre-commitments and creating constraints, and how smart contracts and

blockchain-based systems operate. Because agents in these decentralized systems don't know each other, cannot trust each other, the constraints come through technological guarantees, to ensure others will not defect.

- (Primavera de Filippi) So, as opposed to letting distrust generate more distrust, the question becomes, can we use blockchain technology (or 'trustless systems') as an artifact to create the necessary building blocks for cooperation to emerge, so that those collective, participatory cooperations in turn actually can serve as a basis on to generate more trust?

- (Wessel Reijers, Jessy Kate) Article offers a vacuous, even negative characterization of cooperation. We can aim for a definition that is a bit more thick than simply "not defecting".

- (Morshed Mannan, pointing to the second footnote in the article) Cooperation as agreement of rules, which can emerge implicitly and can also be based on trial and error. A neutral definition.

- (Judith Donath) Game theory does not encompass the variety of human rituals devised to evolve trust by bringing in elements of pleasantness and repeated interactions. Trust can be a way of understanding what it means to be social animals, on top of the coldness of game theoretical models.

- (Judith Donath) Something like bitcoin, and 'trustless systems' remains outside the embodied world of real relationships.

- (Balasz Bodo) Strange to reduce trust to a probabilistic calculus, ignoring the social dimension of life. We don't have such a strict distinction in real life, both elements are constantly overlapping.

- (Andrea Leiter) If smart contracts are a way to replace trust temporarily, leading to something more concrete later on, than what is the type of relationality or societal engagement that would be desired in the horizon?

- (Andrea Leiter) Doubtful about the conceptualization that, like a chemist, we can aim to find the optimal doses of both trust and cooperation so that society functions well.

- (Primavera de Filippi) Less skeptical because consider the article only aims to offer a partial conception of trust, as there isn't one definition that can encompass all of the related processes. A particular contribution that it brings (especially in the context of

blockchain technology) is showing what are the mechanisms that we can use in order to enable cooperation, in a situation where trust is absent.

- (Primavera de Filippi) One path is to increase the constraints (via pre-commitments /smart contracts), the other is to create points of economic incentives and signalling, and that speaks to the case of mining in which you can trust to some extent that the other miners will collaborate because their interests are aligned.

- (Primavera de Filippi) The question is whether the objective is to use this technology to create a trustless system, or to enable a particular environment in which cooperation can be implemented and as result of this initial degree of, even if limited, interaction and cooperation. Can this become a particularly fertile ground where other types of relationships can be developed?

- (Primavera de Filippi) In most blockchains there's the pseudonymity aspect which hinders the process, but in a context where actors in the network use systems of public authentication, perhaps interactions that would never happen otherwise could be enabled by those technological pre-commitments.

- (Nicholas Saul) Back to the question of embodiment of relations of trust and cooperation. Gambetta discusses bonds of friendship and the creation of fictions that allow cooperation to develop. But the text is lacking a further exploration of other aspects of trust and how they are built.

- (Nicholas Saul) The plural subject emerges from a fiction of collective intentionality which redefines how cooperation happens, where it is not limited to the coldness of game theory, for instance as pseudonyms in a blockchain community. The question of whether trust comes before or after collaboration depends on there being an intentionality to cooperate or not. Because if we only face ourselves as one particular individual intentionality than we are missing a lot of different aspects of trust.

- (Georgy Ishmaev) Text is puzzling because it discusses various forms of trust but looks at distrust only in a superficial manner. Distrust doesn't always have to lead to distrust, it's not unidirectional. We can create systems where we can act as if we don't trust each other as a precondition and trust may grow over time.

- (Wessel Reijers) There's an underlying phenomenological, non-explicit point being made: Gambetta's definition of cooperation, is not a definition but more like a theory about how cooperation evolves.

- (Wessel Reijers) The text is missing the element of temporality, which is core to our discussion. What does it mean to trust trust and cooperate, if we can't place these within a specific order of events?

- (Primavera de Filippi) I see the text as surfacing the relevance of these different elements for each other, rather than trying to suggest any particular order.

- (Balasz Bodo) The text warns about not over-fetishizing cooperation, we shouldn't be framing these as necessarily desirable behaviours.

- (Balasz Bodo) Most of the analysis of cooperation assumes that agents have an identity and a reputation. How does the analysis change if we bring in an anonymous environment?

- (Charles Nesson) I resonate strongly with the point that the element of temporality is missing in the discussion.

- (Charles Nesson) Distinguishing between concepts of probability and concepts of belief is a formative question, a central feature of this discussion. Gambetta over-conceptualizes belief in probabilistic terms. The line in which the middle point is a trying point in a progression from uncertainty to belief, and where belief is something that has degrees of probability to it: that's a very powerful game theoretic conceptualization of belief, but it's incomplete.

- (Charles Nesson) To me it comes into question when we add temporality to it. For instance, in a criminal prosecution you start Time 1 with an assertion that someone is guilty, and then you go through a trial where evidence is amassed and the jury concludes beyond reasonable doubt that the crime actually occurred. The question is whether the jury's verdict is a statement of probability or a statement of belief. And ultimately a question of what is meant by "reasonable doubt".

- (Charles Nesson) When it comes around gathering a group as we have, we have some reason to trust each other: we were convened by Primavera, we were chosen. This recommendation is a basis of trust, not a lot of trust but at least it's a place to start. If we imagine that the trust that we have has a temporal quality, and it can grow by acts of cooperation towards the end of building it, I think we can get to the place that Primavera is pointing towards.

- (Primavera de Filippi) If we have more time afterwards I want to continue discussing this distinction between trust versus belief.

- (Quinn DuPont) The question of reputation is really where temporality can be dissected. Gambetta discusses reputation for trustworthiness, which implies something that is built over time.

- (Quinn DuPont) Building a reputation used to be central to the early internet and now blockchain is bringing in the element of trust.

- (Primavera de Filippi) Because of the pseudonymity aspect of blockchain systems, you can't really leverage reputation as a mechanism of cooperation in order to create trust, and attempts to build reputation systems on the blockchain are a response to that.

- (Primavera de Filippi) Gambetta points that you can't prove trustworthiness, only the lack of defection. That's a valuable point in order to argue that it makes sense to trust trust, since otherwise you will never be able to gain more evidence of whether you can actually trust an agent or not.

- (Morshed Mannan) There's interesting research looking at repeated interaction as a path to build a reputation, even in an anonymous environment.

- (Morshed Mannan) The groups that are looking into building reputation systems on the blockchain such as Colony and Daostack, have also been dealing with the question of temporality, but this time not in the sense of accumulation but whether reputation points should be degradable over time. That's a progression over Gambetta's thinking, which can emerge because of the possibilities that this new system affords.

- (Primavera de Filippi) Another way in which blockchain systems deal with the question of reputation is through staking, where you can use assets as a way to signal something about yourself, but that's a strictly game theoretical solution which essentially means these agents are buying their reputation. Because of the way in which the technology is constructed, without formalizing identity, then the system tries to find a way to address these questions through economic and game theoretical means. This means that despite the criticism to Gambetta's paper, it offers a proper representation of how people are actually looking at and understanding trust.

- (Primavera de Filippi) Referring to Luhmann's paper, we can decide to use trust as a way to reduce costs and complexities. I think that that's what Gambetta was trying to express with the language of probability.

- (Georgy Ishmaev) In blockchain the problem is not over how to build trust with an anonymous identity, but the fact that the identities are not persistent. The of how to get people to commit to a single identity remains open.

- (Juan Ortiz Freuler) To me the idea of trusting trust and distrusting distrust brings the question of moving trust from individuals to systems, and then it's important to think about what could be the broader consequences of having digital systems scaling significantly.

- (Primavera de Filippi) Traditional political institutions also shift the focus away from individuals and towards the system, and then we can have more trust in the individuals involved because we know that they are governed by institutions. In a very rough analogy, we could use a technology in a similar way, and ask what are the conditions that need to be there in order to enable this. The question then becomes, what are the societal consequences of having a system in which you don't need to trust people because you can trust the institutions, versus having a people-based trust system.

- (Juan Ortiz Freuler) The value of trust doesn't seem to be limited to cooperation.There's an extra value that emerges when you put trust in someone and cooperation emerges, where you feel like your expectations were fulfilled.

- (Nicholas Saul) There's an interesting footnote where Gambetta discusses how modernity had us abandoning localized similarities and modalities of trust, and allowing for trust relationships to be formed with much less reliable persons.

- (Nicholas Saul) We can use fictions to replace the need for these localized similarities, and that can suffice to establish trust. It works for states because there's a very strong procedure and fiction behind it. Could it work for blockchains? There's an interesting parallel in place.

- (Beatriz Botero) In parallel to these fictions there are always interactions between humans. Even cybernetic places always have chats and places where humans create bonds among themselves.

- (Primavera de Filippi) It's interesting: the institutions shift the focus away from trust, enabling people who would normally not interact with each other to take that additional level of risk, so it ends up enhancing the creation of more trusting relationships by facilitating that initial interaction where you can check the other person's trustworthiness.

- (Primavera de Filippi) On the other hand it can become dangerous. The social credit system in China shows how there's a point in which institutions are so pervasive that the ever possibility of delation makes trusting relationships more difficult to build. So the question becomes, to what extent is an institution supporting an environment where trusting relationships can emerge.

- (Primavera de Filippi) Essentially in the blockchain has crossed that line, where you are so pre-committed, and there are so many constraints that there's no possibility for trust to emerge. Even if an interaction goes well it doesn't give me any feedback on whether the collaborator is a defector or not because the possibility for defection did not exist in the first place.

- (Balasz Bodo) I'm more worried about the commodification and privatization of trust than the institutionalization of trust. New technologies providing trust, including the blockchain, are private infrastructures. They sell trust, this is their product. All the major institutions that have traditionally provided trust have been public, and rightly so, since trust is a public good. But they stand no chance against centralized private providers.

- (Georgy Ishmaev) If we consider trust in a purely instrumental way, as a precondition for cooperation, then we miss the intrinsic value of trust as a part of social life.

- (Primavera de Filippi) Balasz - reputation is a very narrow indicator of trust, so it's easy to commodify. But trust itself is much harder to commodify. What do you mean exactly by commodification of trust?

- (Balasz Bodo) I'm writing about how reputation systems are a way to exert control over the past, automation / smart contracts systems control the present and recommendation systems attempt to control the future. Each of them runs on private infrastructure, each is a market, and there's no public oversight.

- (Charles Nesson) Close the discussion in the form of looking forward or planning for trust rather than trying to build it incrementally in an environment where it doesn't exist. Harvard's central idea is veritas, truth, trust. The idea of establishing an environment for the students where they are capable of developing trusting relationships amongst themselves is central to its mission. Covid showed that the residential aspect was key to enable that.

- (Charles Nesson) The question is, can we approach the blockchain with an explicit intention to build trust, rather than as an incremental and haphazard/vague process. If one

starts from this premise that we want to build trust and that our digital collaboration is valuable, can we use a distributed ledger to build trust around how we structure ourselves?

- (Primavera de Filippi) Do you see any specific function that the blockchain would support?

- (Charles Nesson) The first rule of trust that we engage in is about our understanding of confidentiality. Can I be recorded? If we collaborate in a digital space and then go on to author things that come to have value, this could be exploited. Then in order to properly disclose or distribute each other's contribution, we can have a distributed ledger intermediating a creative commons structure.

- (Primavera de Filippi) One danger of creating additional constraints in an environment where trust is already present is that the transactionality of the processes can actually disrupt trust. There are different costs and benefits.

- (Charles Nesson) I want to invite all of you to join me in playing poker: I mean this in a serious way. Poker is all about messaging conveyed through bets. The game we play does not involve real money. It's education, not gambling, all about getting to know each other and learning about discourse. Georgy and i&i played yesterday after our zoom call ended, talked about the problem of reputation among pseudonymous participants in a cooperating group. Here's the link to my table: all welcome, especially beginners. https://www.pokernow.club/games/fLZdkq4t-__yGbLHS7sLWnm2X
- (Charles Nesson) Even as we parse concepts of confidence and trust we can self-consciously recognize their nuance in relation to one another.
- (Diego Gambetta)

> Protection can nevertheless be a genuine commodity and play a crucial role as a lubricant of economic exchange. In every transaction in which at least one party does not trust the other to comply with the rules, protection becomes desirable, even if it is a poor and costly substitute for trust. This book shows that mafia protection fulfills this role, albeit in an erratic and limited fashion. The market is therefore rational in the sense that there are people who find it in their individual interest to buy mafia protection. While some may be victims of extortion, many others are willing customers. This situation was perceived in the nineteenth century, yet its implications have never been explored in full,

Session III. Feb 11 —Hardin (2002): Trust & Trustworthiness.
Chapter 1: Trust

Attendants:

Key concepts:

Trust, in most cases, is relational: a minimal core part of trust relationships is a clear, fairly well defined interest at stake in the continuation of the relationship.

→ Thus, trust is more than expectations about behavior, or incentive compatibility.

→ *If we have no or only a passing relationship, we are not in a trusting relationship.*

Encapsulated interest theory: I trust you because I think you value the continuation of our relationship, thus it is in your interest to take my interests in the relevant matter seriously (my interests are encapsulated in yours).

This is presenting an account of trust as essentially *rational expectations* about the self-interested behavior of the trusted in a particular matter.

Trust as a three-part relation: A trusts B to do X (there may be a fourth element of context as well)

- Alternative framings see trust as normative or otherwise extrarational, arguing that it is more richly a two-part or even one-part relation than this view implies ("A trusts B", or "A trusts"). However Hardin points out that such alternative senses of trust entail cases that are not always of greatest import in social theory or social life.

Trust as a cognitive notion: in the family of *knowledge, belief,* and *assessment*. Grounded in some sense of what is *true* --- we discover it or are somehow convinced of it (as opposed to choosing it).

→ Assessing competence is a predominant element of trust.

- Certain institutional arrangements arguably eliminate much of the trust that might otherwise have developed between a client and professional, so that our dealings with professionals have more the character of assessing and acting on mere expectations.

Models of encapsulated trust:

One-way trust: where only one of the parties can defect.

Here, if we play the prisoner's dilemma once only, with no expectation of encountering each other again, *it is in the interest of the second mover not to cooperate.* However, this may change if the first mover takes the risk and cooperates -- especially in a scenario where iterated exchanges are possible and the benefits of ongoing cooperation would surpass those of an initial defection.

Mutual trust: reciprocal and grounded in ongoing relationships.

Most common type of trust. Why? Because a good way to get me to be trustworthy in my dealings with you, when you risk acting on your trust of me, is to make me reciprocally depend on your trustworthiness.

→ Several types of behavior often identified as moral can be clearly understood as self-interested in many contexts. Promise keeping, honesty, and fidelity to others often make sense without any presupposition of a distinctively moral commitment beyond interest.

Game theory of mutual trust: commonly, in mutual trust the interaction is a *finitely* iterated exchange or prisoner's dilemma. Standard approach → one should not cooperate in such a game, as final play is a one-shot "defect game". By backwards induction, all plays become defect games.

However, if the first-mover *cooperates* in the first encounter (as Gambetta would put it, rationally trusts trust), the second mover may *reconsider the induction* and decide it is in her interest to reciprocate the cooperation, so that both gain far more than they would from continuous mutual defection.

→ *This ability to reconsider the induction and cooperate is the basis for life in a society, where all relationships are necessarily finite.*

Thick trust: built on thick communal relationships. For Hardin, it is merely a special case of the encapsulated-interest theory, as thick relationships are nothing more than one possible source of *knowledge* for the truster about the trustworthiness of another, and one possible source of *incentives (such as stronger reputation consequences)* to the trusted to be trustworthy.

→ Hardin acknowledges a merit of thick relationship theory: it helps distinguish between individual and institutional problems.

Discussion - Morshed Mannan:

- Hardin's contribution to the already voluminous scholarship on trust is looking at actual rather than just conceptual ideas. In the real world, even in situations that don't lend themselves to trust.

- Main idea in the chapter: trust as encapsulated interest. Accessing temporal interest compatibility is not sufficient for trust. Trust needs to be undergirded by the trustee's interest in continuing the relationship with the truster: "I trust you because I think it is in your interest to attend to my interest in the relevant matter".

- This means trust is both *rational* and *relational*. Trust relationships are limited to certain persons and certain matters - A trusts B to do X.

- Here, trust is a cognitive process, different from Giddens, where trust is less cognitive and more like a leap of faith, and more similar to Luhmann's conception of confidence. Trust as a mental state.

- Whereas trust is a cognitive state/assessment and requires no choice, *acting on trust* entails risk and does require choice.

- Trust can encompass relationships based strictly on self-interest, as well as love.

- One-way trust: where only one of the parties can defect. Mutual trust: distrust can be paralyzing

- Thick relationships: not the primary realm of trust, just one more way of gaining information about others. Concerns about reputation can lead to interest in acting in a trustworthy fashion.

- Trust is primarily rational, essentially based on self-interest, but not necessarily monetary. Can be based on *morality, reciprocity, of self interest* of the trustee. Trust can be maximized by aligning incentives of the trustee.

- Harding lays out three categories of trust relations: iterated interactions or exchanges, which he discussed in this chapter, but also trust relations that are backed by institutions and trust relations that are mediated by non-institutional third parties (discussed later in the book).

- Hardin critiques vagueness of game theoretic work on trust, how it excludes the possibility that players may have a broader relationship.

Relevance for blockchains

- Hardin writes (p.12) that fully programmed automatons cannot be trusted, they are *relied* on.

- Blockchains are not fully programmed automatons. They involve coordination based on meeting predefined and deterministic conditions on-chain, but there's also open ended coordination happening off-chain too. Thus there are elements of both confidence and trust. I'm interested in hearing about whether you agree with this assessment.

- Would the use of a smart contract have prevented the situation that happened with the colonel losing the money transfered illicitly to the merchant, or would the same issue have arised regardless of the use of technology?

Transcript:

- (Juan Ortiz) Towards the end of the piece he says that the types of trust that people have with each other is different from the types of trust people have in institutions. How would this apply to the blockchain? I agree that there are some elements of trust, but wouldn't this system be more like a non-government institution? He discusses trust in institutions in a different chapter, but to what extent are the things in this chapter applicable to the off-chain coordinations?

- (Morshed Mannan) If there's a mapping of the off-chain elements of a particular DAO, conversations in the past years on DAO governance suggest that there's usually a leadership figure that is involved. Most of the time that only becomes clear once a mishap happens, but on other occasions it's something that happens upfront. So there are elements of both institutional trust, but also individual trust, by the person who is interested in becoming a member or participant of the DAO. I don't think that the institutional lens fully applies.

- (Quinn DuPont) Agree that blockchains cannot be fully contained in that automaton definition. There's an interaction between on-chain & off-chain elements.

- (Quinn DuPont) "Generalized trust" is a concept that speaks to people's acceptance of technologies implicitly (high modernity moment). That's part of the formation of institutions as well, so my reading is that generalized trust mitigates much of the need for trust in off-chain coordination. And the other half is just deterministic, so it can be misleading to talk about trust there.

- (Quinn DuPont) We need to be very specific about what is trusted. Trust between cryptocurrency users is widely different from trust in miners. Hardin's text is useful in helping define what we are trusting and how.

- (Morshed Mannan) My question is whether generalized trust should exist? Hardin talks about trust as a cognitive notion. So what is the knowledge building exercise, needed to enable this cognitive notion to emerge? If we are talking about generalized trust on off-chain elements, should this be more nuanced, based on an analysis of whether there are elements of institutional trust, individual trust, and perhaps after that, a degree of generalized trust too. Maybe there needs to be more done to build that knowledge to substantiate this generalized trust. Is it just based on marketing speech, or interactions with individuals involved with the DAO?

- (Nick Couldry) The problem with generalized trust is that when it's broken, it's very hard to know what you have to do to repair it, because it is not based on any articulation of specific elements.

- (Nick Couldry) Luhmann's approach I didn't find useful because it depends on a notion of system, which is very abstract. A great strength of Hardin's approach is this concrete breakdown of trust as encapsulated interest, but it needs to be expanded for our purposes. In Hardin's definition there are very specific people involved: "*I* trust *you* because *I* think it is in *your* interest to attend to *my* interest in the relevant matter". You and I are real,

specific people, not a type of people. And that's the potential weakness of applying this to fit with the blockchain, which is a much more complex and systemic situation.

- (Nick Couldry) I think generalized trust on the other hand is too broad, because it doesn't point to specific layering of things that need to be done in order to build trust and therefore to repair it if something goes wrong. But Hardin's model could help us.

- (Nick Couldry) So staying in the spirit of Hardin's model – which I think is a good one, the best I've seen so far – how can we open it to fit it with the blockchain? Maybe we can turn our attention to what he means by "attend to". Maybe when I use blockchain I attend to the interest of all the other people who are also using it, because I know they must at some level be aligned with my assessment of the risk. Maybe we need to fill this notion of attend to a bit more to get to "alignment".

- (Nick Couldry) Morshed, you summarized encapsulation as *alignment* of interest, but I have a slightly different interpretation on the meaning of alignment. I think that it is a much broader notion than trust encapsulation, because I could be aligned with millions of people for structural reasons, as in the blockchain.

- (Morshed Mannan) Relating this to the example that he uses in the chapter of people driving a car in a highway – where on the one hand trust doesn't exist because you are not committed to the other drives, but on the other hand you must take their interests into account, so there's an incentive for drivers to not crash into each other. I'm curious about what would be the analogy that we could draw from this for something like mining, where they have an interest in the blockchain prospering, and people doing transactions on it, but are not personally committed to each other.

- Primavera. Blockchain systems – because of pseudonymity you cannot have relational elements, all you have is mechanism design and game theory aligning economic interests, and therefore reliability and predictability. Here (as opposed to Gambetta), you might not have to establish trust because the alignment is so strong that it is sufficient.
- Bodo. Text is strange. Sentences sometimes incomprehensive. What he describes as encapsulated interest is everything but trust. The model for the Karamazov example is distrust, not trust. Structuring the incentives of others to enable trust is something closer to coercion than trust.
- Wessel. Found the chapter very confusing. Notion of interest is central yet vague, dk what it means. Temporal aspect is under theorized. Encapsulated interest involves *relationships* & *events*. Both of them refer to interest but what does it mean to have interest?
- Morshed. Backwards induction of distrust → only for that reason do we have living societies.
- Judith. Greatest strength is its emphasis on relationships, but he doesn't take that far enough. The act of being trusted and trusting others in an end in itself (there are interesting biological underpinnings of trust, the experience of trust is pleasant). Karamazov relationships is not affective, purely transactional → the two shouldn't be conflated. With blockchains, to the extent that you eliminate trust, is trust an end in itself that you want to achieve.

- Primavera. Hardin does not analyze trust as an end, he puts it as a means to an end. He identified the fact that I selfishly value the relationship with that person, which is a different value.
- Judith. You may value continuing the relationship, but I'd argue that that's outside the realm of relationships.
- Primavera. So Karamazov' 'example would be outside of trust as encapsulated interest to you?
- Judith. There's an inherent value to being trustworthy.
- Primavera. Agreed that wouldn't consider that merchant in the example to be trustworthy. Maybe a good manipulative human being.
- Philemon. If trust is just based on interest than its confidence. Thick trust is what I would define as trust. Hardin's definition is too economic, transactional.
- Primavera. Perhaps there'' a distinction to be made: do you only trust because you think that the person is trustworthy or can you also trust because the trustee has encapsulated interest. That's why Hardin frames it as a three-part relationship, because I can trust someone in a specific set, but not in others.
- Balazs. If you go to a restaurant you can trust that it is complying with food safety rules, and I trust his self interest in complying.
- Morshed. A third party institution is what enables that trust. In the absence of it, you may consider reputational effects, but maybe the self interest of the restaurant could be to maximize profits.
- Primavera. Independent of those two, if I am confident that the restaurant wants me to come back as a customer that I can trust that the restaurant has my encapsulated interest.
- Balazs Bodo. In modern societies it's rare that we develop relationships where you are seen as a human, not just a client (your bartender knows your name and wants to make sure you are satisfied).
- Judith. Aspect of modernity is that you can
- Charles. Is trust possible without human trust? Could a smart contract have prevented the lieutenant losing his money?
- Primavera. Does not solve the problem to trust, hampers it because the very fact that you are using the smart contract implies you have no trust.
- Charles. Isn't the premise to use blockchains as a base to lead to trusting relationships?
- Primavera. Larger quest is to try to avoid the problem of reconstructing trust and just leverage alternative means to enable cooperation. My quest is to try to understand how new types of relationships enabled by the blockchain can bootstrap the basis for trust.
- Charles. Emotional state is a product of senses. Cognitive state is a product of the brain. When I think of trust I'm in an emotional state. Cognitive state, is math and logic, and very trustworthy in most senses, but has no ethics.
- Georgy. Hardin is very critical of general trust, which is what we have with banks, and facebook. He tried to introduce rational trust, which is a better framework for cooperation.
- Wessel. Agree trust is emotion. I also think it is a virtue, which is emotional but rationally mediated. My problem with framing it rationally is that it becomes a worldly affair. Virtue is in between rationality and emotion. Trusting itself as a virtue. Anticipatory resoluteness as a definition of trust.

- Primavera. Trust as encapsulated interest can come from values, but there can be multiple grounds, value-based grounds, cognitive grounds, etc.
- Paula. If trust can be understood as Gambetta puts it: a line from complete distrust to complete trust with a mid-point of uncertainty, so numerically ranging from -1 to 1, then in a particular circumstance where A trusts B to do X, can a smart contract be seen as increasing the probability for trust to emerge as opposed to uncertainty or mistrust? Just by the mere fact that it obliterates the range of -1 to -0,1. In this framework, even when everything goes according to what has been programmed, trust can remain on 0 - you have no information about your counterpart's trustworthiness. But I am interested in the question of whether we can affirm that the probability for trust to emerge, as opposed to mistrust, increases? (of course, not in Luhmman's sense, who would just frame the interaction as one of confidence).

March 2 —Nissenbaum (2004): Trust as Security (Guest speaker)

Attendants: Helen Nissenbaum, Quinn Dupont, Divya Siddarth, Philemon Poux, Morshed Mannan, Balasz Bodo, Wessel Reijers, Georgy Ishmaev, Primavera de Filippi, Aviv Barnoy, Heleen Janssen,

Transcript:

- (Helen Nissembaum) When I was writing this paper, the National Scientific Foundation had a program (still has), where computer scientists were merging ideas of security and trust. I was trying to argue that it's not ok to call something "trust" if your aim is security.
- (Helen Nissembaum) Re-reading the paper - it's depressing how the "insider" and "outsider" part remains relevant. The main threat comes from respectable parties, and we have to engage with them. Things have gotten pretty dire in that respect.
- (Balász Bodó) The most interesting thing is how the paper has aged.

Key concepts & discussion (Balász Bodó)

- *The argument:* The assumption is that we need trust because lack of trust is paralyzing. That was especially relevant in the context of e-commerce, with frauds. To have a lively digital economy you need trust.
- *Trust is caused by:* history and reputation; inferences on characteristics; mutuality and reciprocity; role fulfillment and contextual factors (cheating can be made public / rewarded / punished). But many of these factors are not present online.

- *Problems of trust online: m*issing identities (in the internet nobody knows you are a dog), inscrutable context (digital domain is a context of its own, disembodied from institutions that facilitate personal relationships in real life.).
- *Solutions that have been proposed in many technical discussions:* trust through security, better access control, management and enforcement of identity and surveillance
- **Argument of the paper: these are misguided approaches. Trust depends on the freedom of others. Certainty, security and safety, give a certain type of warranty but comes at the cost of freedom.**
- Trust in information systems is the question of insiders / outsiders. *Facebook is not an information system in itself, it's a firm that has its own agency*. Can we trust it as a firm (as opposed to a technical system)?
- More insidious setup now: these systems grew into monopolies, and we are now stuck with untrustworthy systems.
- The question is not about whether you use them or not (we are stuck with these options) but how do you live with untrustworthy systems?

The threat model changed:

- Missing identities is not the biggest problem anymore - we know too much about each other and platforms know too much about us.
- Evil hackers are not the only or most important enemies, there are many other problems: greedy corporations, extractive business models, state adversaries, unaccountable CEOs, lack of regulation, unexplainable ML models

Problems with the solution

- Sometimes we want distrust. The whole system of democratic institutions comes from carefully managed distrust.
- Sometimes we want security. We want to have some degrees of certainty in how we relate to apps and platforms.
- False dichotomies: trust enables people to interact in a rich and complex world, whereas security decreases richness and complexity. Trust and Security are not necessarily mutually exclusive. We are in a richer and more complex digital world, and we need to find a place for both trust and security.

--

- (Helen Nissembaum) Minor side comment: I realized that back then there was the issue of identity - those were the days of nobody knows you are a dog. The interesting point is that we are talking about the factors that seem to lead people to trust others, and identity was one. But what often is the case, is that identity can be pseudonymous, but knowing that you will have a continuous relationship with the other is what enables trust.
- (Helen Nissembaum) The problem with algorithmic governance is that there's no accountability. It's not about identification, but the acceptance of action without accountability which emerges as the problem.
- (Balász Bodó) With the big platforms - social media, airbnb - their real product is reputation. But not accountability, because it's so easy to re-register and create false profiles.
- (Balász Bodó) Blockchain systems were also developed to respond to the anonymity problem. It's a completely secure and safe system, there is no freedom there. You don't choose the rules, you only have the freedom to join or not.
- (Primavera de Filippi) I find your paper timely. At the time, the fear was how do we protect ourselves against hackers. And in large part this has been achieved, we do feel safe to pay things with our credit cards. And yet, we never had less trust in online operators than today - we are not in a realm of trust. So the external security aspect has been dealt with, but not the internal security: technological guarantees, accountability, separation of power, within the system itself.
- (Primavera de Filippi) When you increase security too much, you don't leave room for trust. But our problems today, are they due to excessive security or lack of security against insiders?
- (Helen Nissembaum) In re-reading the paper, I kept thinking to myself - so what? If we make things so safe why do we care about trust? Maybe it's better to not have trust. Maybe we should celebrate the security we do have.
- (Helen Nissembaum) In the online realm, I speculate that we secured certain parts, that's great. We still have to trust the internal people - like your mobile operating system. And yet, we don't have the information or context or all those factors that I listed, about those folks in order to wholeheartedly trust them. And these guys are con artists - you can't trust them more than a used car salesman. We don't have the regulation that would require the needed auditing.
- (Helen Nissembaum) But while you can be secure against certain types of aggression or attacks, you cannot live your life without trust because you will always be vulnerable.

- (Wessel Reijers) What struck me was how trust is built through a specific setting of identity and reputation, which is very similar to Chinese Social Credit system. How does a trust based system, understood along the lines of the context that you sketch, improves freedom? Chinese Social Credit system has all of these criteria but does not increase freedom. Similar with blockchains.
- (Wessel Reijers) The ideal of trust as security is false. You only have walled gardens. Question becomes: is there a path that increases freedom?
- (Helen Nissembaum) What is it about blockchains that makes it a libertarian favorite, in what way does it enhance freedom or liberty?
- (Quinn DuPont) Blockchains don't promote freedom in the sense of, do whatever you want. They are a technique against the state.
- (Primavera de Filippi) Blockchains enable you to bypass certain institutions that liberals don't want to be subjected to. You escape from one type of constraints, like regulations, and then you choose your own set of constraints, among the different blockchains. So it's not like "I want to be free of everything" but "I want to be able to choose my constraints".
- (Helen Nissembaum) The fact that libertarians are so committed to the preservation of property rights, and the way they impose on the freedom of others in doing so strikes me as odd. Some limitations on freedom are blessed by them, while others are not.
- (Morshed Mannan) At some stage blockchain libertarians run into some regulations. In Kenya they passed legislation to allow for blockchain free-trade zones: so people realize that there are certain constraints outside those zones. A regulatory sandbox in a geographic space. This is interesting to question the claim of how far blockchains are free from regulatory constraints.
- (Helen Nissembaum) Why, as people studying blockchain, are you studying trust?
- (Primavera de Filippi) In the blockchain system there's no insider or outsiders. You don't trust anyone, so you need technological guarantees or economic incentives to ensure that actors will act as expected.
- (Primavera de Filippi) On the other hand you have institutions analyzing the possibility of adopting blockchain technologies. So our question is, can we actually use blockchain technology to increase insider security? And by increasing insider security and confidence, could that contribute to increased trust. Or if we secure it too much we may eliminate it? Are there fields where we can use it?
- (Helen Nissembaum) Assuming we could secure more, what will we lose? What relationships are the most degraded ones?

- (Heleen Janssen) Can we discuss gradations in trust and security? Distrust can be useful, it can keep people communicating with each other. You need breathing space and dialogue.
- (Helen Nissembaum) Technology constraints can shape behavior and create its own morality. But then that's like the shackled psychopath, it's only the shackles that are holding him back. With The DAO hack, is the blockchain implementing a moral rule, or is it the moral rule?
- (Balász Bodó) Attractive dichotomy, but in practice there are very few situations where this comes out very clearly. We don't want to trust that the uber drive will not rape us, we want to have confidence.
- (Balász Bodó) Digital world: most of the trust producing institutions are private (offline, most of them are public at least in the modern era).
- (Balász Bodó) With technical systems, the problem is not that they are producing too much security and eliminating trust, but that they are commodifying it. Safety is being produced by private contractors.

March 11 —Bodo (2020): Mediated Trust (Guest speaker)

Attendants:

Transcript:

Introduction by Andrea Leiter

- Paper complements existing theoretical thinking on trust with the question of what does it mean when technology becomes The Entity that is entrusted to be the mediator of trust.
- Trust is conceptualized both as an interpersonal relationship, and an institutional relationship. However, both on the interpersonal and the institutional level, the trust relationship is characterized as an interhuman and intersubjective relationship. The paper is genuinely concerned with how humans trust each other.
- Crisis of trust - perpetuated through two things that are taking place at once:

    - Globalization: The nation state, as the one entity that in a sense bounded trust, does not scale up to the global level.
    - Digitalization: Digital technologies now create this new phenomenon of technologically mediated trust.
- Three questions the paper asks:
    - How do digital technologies establish new forms of interpersonal and institutional trust?
    - How digital technologies transform the existing logics of interpersonal and institutional trust?
    - And ultimately, how can we establish the trustworthiness of trust mediating technologies?
        - How can we assess this transformation? What kind of tools do we have to see what is coming for us? What game are we in? Is that a good thing or bad thing? What do we do with this?
- ABI framework: trustworthiness of technologies fail the parameters of Ability Benevolence and Integrity.

<u>Discussion by Andrea Leiter</u>

- When do we start having technologically mediated trust that we did not have before?
    - What is the fantasy of unmediated trust that stands on the other side of a mediated trust?
    - How would unmediated trust look like? In a sense trust is mediated through all sorts of technologies be it, the telephone, the book, the paper and even on a more fundamental level, words. So the question becomes: what is so fundamentally different about these digital technologies that they really are not just part of the story, but you want to ascribe agency to them in a new quality.
- It seems that what changed the notion of trust, is the moment it became a response to risk.
- You speak of how technologies actively shape our relationship to trust. Going away from only the human being the subject, and everything else being is an object; towards seeing an agentive relationship where there is a co-production at stake.
- On the heels of this co-production, one thing that I would provoke you with is to say that the lens of risk in and of itself is one that is co-produced through the tools of governance. Mapping, cartography, statistics and everything that we use in order to assess

the world and measure the world has now brought us into this category of coding, numbers and symbols, assessing them on a measure of *risk.*

- So risk is the category that makes us engage in the world. But I wouldn't say that risk is the only category that makes us engage in the world and always has been. Risk only came to be invented as that category through the tools that we deploy to make sense of the world. And so that trust in a sense becomes only a response to that. The notion of having to use trust as a risk responding mechanism to engage in this world is already embedded when we ascribe the risk category as the one with which we want to make sense of this world.

- Thinking of trust as a response to risk appears as a necessity, if we ascribe to the tools and to their mode of operation. That is already something that we might want to challenge and think through. Is that a good way?

- Other than saying are these tools trustworthy? We might want to ask: Does it make sense to think through the lens of risk? Is that the lens through which we want to know our world and through which we can engage in the world?

- The final provocation I would like to put out is on human reasoning and reasoning in general, that to me appears to come through as the thing that has to be the safeguard.

- In the assessment of ABI, what is lacking in a sense is transparency and an idea of comprehending the logic that is being displayed by these technologies in order to domesticate them, to make them graspable.

- And I think that algorithmic reasoning or machine learning reasoning escapes this idea of domestication, or bringing it back to human reasoning and rationality, making it penetrable with our knowledge. I don't think that is possible.

- Here I was very much taken by Louise Amoore's book Cloud Ethics, in which she has an approach that is an embracing of doubt and the unknowability. She tries to come to terms with technological reasoning or algorithmic reasoning, not by trying to keep understanding it with the human mind, to open the black box, make it transparent, know the reasoning and the causalities and follow it back; but by going at it in a different way that says, this is an instantaneous decision making process, and it is a way of governance that operates differently than the logic and planning that we are that we are used to so far.

- And so I think that there could be something in that direction of trying to think with the unknowability, and try to capture it in a different way in order to make it trustworthy.

- (Balázs Bodó) So about "when does the new form of mediation start?". This whole paper is based on two fundamental texts. One is Susan Shapiro's The Social Control of Impersonal Trust, and the other is Lynne G. Zucker, Production of Trust: Institutional

Sources of Economic Structure from 1840 to 90. Zucker's article is a description of a big transformation of the US economy with the settler culture that comes in. She does a meticulous documentation of the economic activities there, describing how an economy based on knowledge, familiarity and customs has to give way to something else because it can't scale with the new settlers. So a new way of establishing trust has to be built. The technologies didn't change but the socioeconomic realities changed, so new ways of building trust were necessary. When we are dealing with unfamiliar faces on alibaba, or when the 2008 crisis in the USA makes my life difficult in Hungary, there's a need to mediate trust in a new way that applies to these new circumstances.

- (Balázs Bodó) If reputation and repeated interactions are key to trust, my question is how do we remember these interactions? It depends on the medium, and whenever there's a new medium, the type of trust that is built changes. Different mediums have different ways of operating. When social changes happen, we also develop new ways to establish trust with a medium that is more adequate to the new reality.

- (Balázs Bodó) So my answer to, when this new form of mediation starts, would be in the 2000s when we see literature emerging on how to make e-commerce more trustworthy. I would say that's the beginning.

- (Balázs Bodó) The second question is whether the lens of risk is the one through which we want to engage with the world? I would be very interested in discussing what other lens can we use to engage with the world. How else?

- (Balázs Bodó) The last problem is the domestication of technology - how can we domesticate systems that are intransparent, or very speculative or deterministic. We had a similar problem: how do we domesticate secret services? Institutions which are by definition intransparent, and resist oversight. If something like machine learning models are inscrutable for human oversight, if that is what we have to work with, then maybe we can apply the logic devoted to oversee these organizations.

- (Balázs Bodó) You mentioned Cloud Ethics, I can offer another book on the topic, Technologies of Speculation, which also speaks about speculative futures, and what happens when we can only speculate but can't reach any certainty, and how we deal with that type of uncertainty.

- (Primavera de Filippi) I was expecting to see more in the session where you describe the tools being used to facilitate the establishment of trust. One based on the past, reputation. One based on the present, blockchain. And one based on the future, recommendation systems. I think it's a very nice three-part separation. And I actually want to open the question for discussion, of whether either of those three are actually successful, effective tools to establish trust? Or whether they reduce the need for trust. To me blockchain is not a technology that actually creates or enables new trust, but reduces the need for it. So

to which extent do you see these as establishing trust, or tools that avoid the need for trust intermediaries?

- (Balázs Bodó) I think this is a discussion that has been going on for quite some time between us. You are focused on trust and confidence. A differentiation of a reliance on some form of control or of confidence in the actions of others, and something that precedes that and is more an expectation for myself to deal with the uncertainty and the lack of control.

- (Balázs Bodó) I don't really differentiate between trust and confidence. I think in every situation in life they are mixed together. Some part trust, some part confidence, some part blind faith.

- (Balázs Bodó) Reputation is an important signal for e-commerce - and that's aggregating the past right? The past transactions of both parties. Is it confidence? Is it trust? Is it faith? I think it's all that. But that requires identity. And what I see with the blockchain systems is that they prepared the solution where there are no fixed identities, which enables transactions in an anonymous environment. That comes with control, right? This is how they control the present, blockchain systems, smart contract systems create certainty in your vocabulary, by creating a certain type of certainty to control the actions of the counterpart.

- (Balázs Bodó) So it would be interesting to actually measure empirically: to what extent this is a confidence based on the full understanding of how the system works and how much is it just because someone told me the system is reliable. For me the DAO hack is the case in which the limits of confidence become very clear. People were very confident because this was open source code, but there was an element of speculation as well which goes beyond confidence.

- (Balázs Bodó) And with AI and recommendation systems, my argument is that they build trust by being able to reduce future uncertainty. If we can reduce future uncertainties, probably we need less of this mix of trust, confidence and faith. So I think they can play a really important role if we believe that they will lead us to solid ground. These systems are reducing uncertainty but not increasing certainty, and this is where the trustworthiness question comes from. Are they actually delivering, or are they increasing the risks?

- (Primavera de Filippi) Maybe this relates to the question that Andrea raised about risk. Which is that perhaps trust cannot be only about reducing risk, uncertainty or vulnerability. It also requires some type of choice and agency, capacity to act. And the interesting thing about this section on the three approaches is they all reduce risk but not all of them have the same impact on choice and agency.

- (Primavera de Filippi) Reputation reduces risk because I get more information about who I'm interacting with, therefore it increases my agency of choosing: do I want to interact with that actor, given the trust behavior of this person? It's providing more information, hence enabling me to expand my realm of action. I can make more informed choices.

- (Primavera de Filippi) The blockchain, I define more as a confidence setting, because it does reduce the risk, but it does not enable me to even question anything. I don't need to trust because there is no question that it's going to operate as expected. And of course there are all the drawbacks of that, because confidence is not certainty, so there can be problems. But the blockchain is basically trying to prevent the system from doing anything outside of what has been codified into the protocol. So it's reducing risk, but also reducing agency.

- (Primavera de Filippi) And with recommendation systems, it seems we are going even further in the reduction of agency, in the sense that it's not just preventing me from doing things that the system does not allow; but it's also pushing me to do what the system thinks I should do. And so it's even further reducing my agency. It's nudging and suggesting. Maybe it's not even confidence anymore, maybe its a third thing. While I see reputation as a trust building mechanism, and I see blockchain as a confidence building machine, I'm wondering whether recommendation can actually qualify as trust-building, or even confidence-building. Or is it a completely different type of risk-minimization. But that also comes along with a very strong impact on individual agency. And can we even talk about trust-building or is it something else?

- (Wessel Reijers) I have a question about the section on institutions of trust production and distrust management. Which is a bit too short because there's lots in there and there was something I actually didn't understand. So this achieving trust through distrust -- I didn't really understand how we're going to get that. "*Achieving trust through distrust entails establishing systems of accountability, checks and balances, oversight and supervision, backup systems, and insurance, which disincentivize, detect, punish, and remedy the breach of trust by these institutions.*"

- (Wessel Reijers) So at first actually I had a different kind of notion in mind. Which is a bit more like Cold War logic as in like through the cultivation of the distrust. So in the Cold War you have these two parties that severely distrust each other, and through the distrust they are incentivized to arm up, so you get an arms race. But because there's a threshold that none of the parties wants to cross -- paradoxically this arms race creates trust through distrust. That's how I understood it first.

- (Wessel Reijers) But then I read that you discussed this notion of accountability and checks and balances and so forth. So I wonder how do these two understandings of trust through distrust relate to each other.

- (Wessel Reijers) My other question has to do with the last paragraph: "*These different logics produce a spectrum of trust. At one end of this spectrum, risks, contingencies, or even fate (or fortune) is managed through faith: a non-cognitive, non- verifiable belief in some positive outcome.* " I find it an interesting statement, because *managing*, as in acting on something knowable and believable, and *faith*, which falls outside of these categories, together in one sentence seems hard. So that would be my second question: how should we understand that apparent contradiction.

- (Balász Bodó) I was thinking about how the imaginaries of technology are very similar to how the Catholic faith is managed. It is not an accident that every Silicon Valley company has a chief evangelization officer. Trust in digital technology is very much rested in this unverifiable, unjustified, irrational belief that technology will make us better or make society better. Its evangelization. And not by accident it is called this way, because at this stage there is little more to actually support these ideological claims and political imaginaries that came from the 60s counterculture, or libertarian. corners of the Internet. And if you look at these cult figures like the Apple cult and blockchain cult, its very similar. Lots of pseudo religion.

- (Balász Bodó) To your other question, I added the link of Sztompka' Trust, Distrust and the Two Paradoxes of Democracy. You are very right in your observation that mutually assured destruction is one way of building trust through distrust. It's pure game theory. What Sztompka analyzed is something else: in democracy, you trust democratic institutions, but trust them by not trusting them. So we trust politicians, because every four years we can get rid of them and we trust elections only because we know that there are international observers to make sure that the elections are good. You have state bodies who actually manage elections and we know that they are under the supervision of elected representatives. So there is this institutional method where everybody's watching each other. And I can be reasonably confident in the trustworthiness of an institution because I know that it is embedded in a system of distrust.

- (Balász Bodó) And I see that this is what has collapsed in the 2008 financial crisis. We thought that the banks were trustworthy because they had management to make sure that they were trustworthy, but they were not. We know that boards oversee the management, but the boards were not doing their job. We have credit rating agencies to make sure the products of the banks are audited, and they were not doing their job. We thought the financial authorities and elected politicians were doing their job, but they were not doing their jobs. So what happened in 2008 was not just the collapse of a few banks or mortgage institutions, but this whole system of trust. This system of distrust

collapsed. And that's not necessarily the same thing that a prisoners dilemma type of situation.

- (Quinn Dupont) The way that you work through the problematics striked me as getting away from this sharply liberal focus on trust, which I think could permeate a lot of our discussion thus far and I think it permeates a lot of the general discussion about trust. We kind of take it as this sort of very liberally central centered focus right. When it comes to questions of responsibility and power and ethics and we tend to elevate considerations that, I think, miss the boat. For instance privacy becomes really important if you think of some agent being at the center of it all. But you don't look very deeply to understand that the ills of the Internet have little to do with privacy and much more to do with other kinds of support.

- (Quinn Dupont) ABI doesn't match how my brain works. It seems to go orthogonal in some sense to how my brain thinks. So you've got benevolence: this is a directly ethical category. But interestingly, you pick integrity and say this the ethical category. And I'm just reminded of the Belmont Report or the (?) Report, which are these two UN-based guidelines for how to do research on ethics. They have categories of guidelines for ethical assessment, and those include respect for law and public interest.

- (Quinn Dupont) One way of reading the word integrity is, not being a hypocrite or something. As the kind of ethical framing of this. So I'd love to know more about where the ethics of these systems are located.

- (Balász Bodó) I've noticed how difficult it is to actually maintain our ethical rules or norms of engagement or integrity, in a system that spans multiple cultures. This is very similar to the challenge we had at the age of the railroad, where there was a question of how do we control the technology that spans multiple time zones. And now it's a question of how we control a technology which spans multiple contradictory ethical systems or cultural logics. The same technology or the same firm that produces the technology may want to do business with authoritarian China, libertarian US, human rights actors. And these feelings are transparent to each mediums or locales. Then you just have to start dealing with employees saying we are out, we are on strike.

- (Balász Bodó) This is also relevant for the blockchain space. On the one hand there is all this talk about how beneficial these systems can be. And then there is this counter-discourse saying there are environmental costs, an ethical claim about the integrity of a resource-hungry technology. NFTs are being mined at exorbitant energy costs.

- (Balász Bodó) And blockchains are dispersed throughout space, they are not isolated. It's not something happening in my backyard, it's not an individual decision, and this aspect I find fascinating.

- (Quinn Dupont) I'm interested in how benevolence and integrity are ethical categories that map in non-standard ways to existing discussions on ethics and power. My question is what are our responsibilities to trust systems? The development of them, the use of them.

- (Balász Bodó) This discussion of benevolence: does the trustee encapsulate, or acts in my best interests? That's benevolence. Within the trust relationship there is this encapsulated interest of mine, right? This discussion comes up when a great actor turns out to be a sexual predator or a great philosopher turns out to be a nazi, or a great doctor turns out to be hunting an endangered species. In their capacity of entertainer, or in their capacity of doctor, or lawyer, their benevolence towards me can be demonstrated. They act in my best interest. But they are monsters by my values system. Can I watch the films of X after learning they are sexual predator? There are no good answers for this - or maybe there are, but I see significant confusion about these issues. Those are the dimensions of benevolence, integrity and ethics where there can be potential conflict.
- (Andrea Leiter) Primavera, based on your agency point that you made at the beginning. One of the things that I was thinking of when I was reading your paper Balázs was the notion of force, of disciplining. So if the positive way of going about it, would be agency, a different formulation of it would be that of resistance, rebellion, push back. So if you think of what kind of engagements are possible on the blockchain, what kind of place of resistance do you have against these mechanisms? Not too many. It's pretty straightforward, you either take it or leave it, and you either play by the rules or you don't. So this agency limitation that you mentioned Primavera, also translates into the notion of force that comes with trust. It can be enforced. There is a force and a violence, a disciplining into a particular kind of behavior.

- (Andrea Leiter) And Quinn, what I heard from you that maps on to me is: where is the responsibility for that? What is the disciplining body here? In a sense it's the structure, the architecture itself. An entanglement of different things that is doing the disciplining. And that's a very different way. From my legal background, it's a question of allocation of responsibility. In the liberal framework: here is the subject, here's the acting agent to whom I allocate responsibility for. But where do you allocate responsibility for a disciplining architecture?

- (Primavera de Filippi) I like this comment a lot Andrea, and I think now I managed to make sense of what Quinn was referring to, which is that with trust, when we discuss it

on an individual level the ethical question is important only to the extent that I'm evaluating a trustee. But there's no external or ecosystemic need for an ethical framework. Whereas the paper is pointing out how trust mediators are not just mediating trust, but they are also indirectly changing the way in which we behave, completely changing how we operate. And when I use a blockchain or a social media with extreme customization and recommendations, because it is affecting my capacity of action, then in that case there comes an ethical framework that needs to be imposed on these actors. And it's not because I decide to trust them, but because by the mere act of using them, whether it is driven by trust, confidence or anything else, it's having an effect on the way I behave and interact with others. And then the ethical question becomes fundamental. Is that what you were saying Quinn?

- (Quinn Dupont) The virtue of this article is that it gives us some purchase on these other kinds of framings, even though the paper doesn't really talk about it overtly, I think it helps us change the framing in the way that you kind of pointed out.

- (Primavera de Filippi) I like the language the paper used to frame these contextual things. There is interpersonal trust, but then thats mediated by an external system which can be a specific institution. But then of course why would we trust those trust mediators? It is because there are these larger institutions, which are also creating the necessary trust or confidence for us to trust the trust mediator. So you just keep expanding: then why do you trust those larger institutions? Well because you have even larger institutions like governments, which also all contributes to creating a system in which it is ok to rely on those trust mediators.

- (Primavera de Filippi) The trust mediator by itself cannot self-fulfill the mission of establishing trust, because you need to trust them. And the same applies to blockchain, you have confidence in the technology but you only have confidence because you trust the underlying governance system. Why do you trust the underlying governance system? Because you trust the larger economic incentives, etc. So I don't know if it's explicitly stated or if it's just something that transpired from the reading, but to me I really liked this conclusion of the paper that indeed the problem today is that with those trust mediators, there isn't sufficient justification coming from external systems for us to trust them. And maybe the problem is not with the trust mediator itself, but its about the lack of an external framework that can make sure that those trust mediators are trustworthy, and therefore we can't rely on them to trust at the interpersonal level.

- (Balász Bodó) Or build interfaces. I really like descriptions of trading methods because trade is built on trust. And the question is, in ancient times, in medieval Europe or the Silk Road trade, what is the framework in which people are able to trust each other to

carry goods for them, or carry money for them, and build a continental wide trading system, or multiple trading systems. It spanned from Vladivostok through Africa. And it's not by accident that the most of these trading networks are organized by religion, or by ethnicity. Armenians traders, jewish traders, arabic traders, the trust framework in which they operate is religion. It's Islam, Judaism, the Christian faith, the same ethnicity, the same language. These are the trust frameworks.

- (Balász Bodó) And the question is, at what point can they interface with each other, so a Jewish trader can interface with an Arab trader. And what is the city as another frame where these two systems meet. And where these two systems can engage with each other despite the fact that they are not given the same trust systems. Maybe for blockchain systems its enough that its a self-contained trust system, like the Armenian trade networks, which have their own trading outposts in Amsterdam, in London, everywhere.

March 24th: Schneier, B. (2012). Liars & outliers, Chapter 6 (Guest Speaker)

Attendants: Bruce Schneier, Quinn Dupont, Primavera de Filippi, Matt Prewitt, Aviv Barnoy, Shawn Bayer, Wessel Reijers, Balász Bodó, Georgy Ishmaev, Morshed Mannan, Juan Ortiz Feuler, Andrea Leiter, Lana Swartz, Eric Alston, Philémon Poux, Nicholas Saul, Viet Ho, Judith Donath, Charles Nesson, Paula Berman.

Transcript:

Introduction by Quinn Dupont

PDF

- I should just start with talking a little bit about the interesting task we have in front of us. We're blessed to have the author Bruce Schneier right here with us, so it makes my task slightly interesting, and different.

- (Bruce Schneier) I wrote it 10 years ago, so you probably know it better than I do.

- Quinn - When we read things we're looking to kind of grab some theoretical apparatuses or some interesting examples that we can use on our own research. But I didn't really think that made a whole lot of sense necessarily. And then another approach that we've often taken has been to sort of say, "we're reading Luhmann, it was rather unclear exactly what he is talking about". But the thing is, Bruce's book is beautifully written and it's actually really clear. I read most of the book. And it's the kind of book that I wish I could have written.

- So I'm going to approach this with a little bit of exegesis first. And then I have a slide where I think: if I was to have written this book, maybe what I would have done a little bit differently.

- Um, we started chapter six on societal pressures and we encountered these societal dilemmas, also known as these coordination games, right?

- And our societal dilemmas are basically choices between group interests and some competing individual interests.

- But in the real world, it's more complicated than that. And Bruce does a magnificent job of really bringing us back to empirical examples that are originally contextualized in the real world.

- So the chapter talks about societal pressures, and that's to say making it in people's best interests to act in group's interest. And there's a wide variety of mechanisms that constitutes the societal pressure that enable cooperation, but the defection happens anyways. So there's fundamental tension.

- And I think one of the goals is cooperation. Bruce writes division of labor is an exercise of trust, and trust seems to me to be epiphenomenal. In chapter seven, there's examples like: your one vote never really matters, but democracy only really works when everybody votes.

- So on mechanisms of societal pressure, the first is this idea that we can modify the game. The chapter extends that a little further and unpacks it, in ways that will hopefully generate discussion when we talk about blockchain technologies. To modify the game, that's a form of reverse game theory, which is also known by the name of mechanism design, also known by the name of crypto economics.

- And then we go into the typology of these social pressures. There's moral pressures. Reputational pressures, institutional pressure and security systems. And they are put

together as mechanisms for coordination. So there's an increasing of the difficulty of defecting, maybe with consequences, reducing benefits of defecting, limiting the damage caused by the faction, increasing the benefits of cooperation and lowering the cost of cooperation.

- A lot of that is game theory, you know, change that mechanism, modify that game and get the desirable outcome. Okay. So very briefly, I will give a quick snapshot of the chapters that outline each one of these societal pressures.

- **Moral pressure:** As I read it, there's an ambiguity here, on either societal or individual foundations for that. Bruce writes "any innate or cultural guidelines". It's kind of, uh, I think in many respects are akin to psychological traits or tendencies. Somewhat provocatively, there's this idea of natural or cultural evolution, which produces moral receptors.

- Of course, moral pressure includes rewards and penalties, and as well being moral pressures, they need not be formal. That's in distinction to institutional pressures. And more pressure is great really for small groups. And in fact, it's hard to enforce in large groups, especially when we have arbitrary, moral pressures, right? This kind of tension you get when there's an authoritarian regime, that puts a moral pressure that doesn't start naturally.

- **Reputational pressures:** these work best in small to medium groups. We're not really great at detecting defection, especially when it comes to larger groups. And of course also, when you talk about the internet, when there's geographical distances. Reputation used to be largely local. Now things are a little different.

- And both the moral pressure or reputational pressure is constrained by a number of psychological factors. A lot of examples that Bruce uses: in-group preferences, Dunbar's number, et cetera. And these are social psychological factors and they need consequences to be effective. That's one of the things that reputational pressure kind of needs to really get going.

- **Institutional pressure:** this works best in large groups, and it can be used to solve the tragedy of the unmanaged commons. And of course, there's also Ostrom's rules for polycentric governance, which were all familiar with, and it's found here in Bruce's book as well. Institutional pressures can be effectively formalized. Bruce drops a very solid line between institutional pressure and his conception of law. However, they don't always accomplish these goals. So the example here is that fines not necessarily to deter speeders, or corporations will take their fines and just roll that into part of their operating budget.

Right? So these are some of the challenges where institutional pressure breaks down. And then ultimately institutional pressure can be really socially expensive, and that's presumably not a good thing.

- And then we come to security systems. On the one hand, Bruce says, these are everywhere, right? We have a really rich wide conception of security systems, as a kind of weird hybrid, right? They work across many scales and time periods in ways that other pressures don't necessarily work quite as effectively. Security systems are able to fill in the gaps from where other sort of societal pressures don't quite manage to do what they need to do.

- Security systems are envisioned as an analog to our natural defenses, and really interestingly they're the only societal pressure that puts actual physical constraints on behavior. I think this is a really important idea to think about especially when we talk about the blockchain. We can use security systems to augment any of the other societal pressures, so we can have security augmented, or enhanced, moral or reputational or institutional forms of pressure. So you can kind of mix and match, add a little security, but too much security ends up being a problem. Both in the implementation, but also there seems to be a bit of a moral claim going on here as well. It's not entirely clear to me exactly what that moral claim is, but I see an undercurrent.

Discussion by Quinn Dupont

- This is my last slide where I present: if I was to have written this book what I would've done a little differently?

- The first thing to note is that Bruce is really clear that societal pressures and dilemmas aren't intended to be reductive. I think there's one uncharitable way of reading the book is to say that, "Oh, well, you're just talking about, um, sociobiology" and I don't think that's a fair characterization of what's going on here.

- Definitely, we've got game theory and rational agents. We've got psychological predispositions, we've got morality, we've got organizational arrangement, but I take it to be kind of an invocation of a pluralism. And one of the really nice things about this book is that it's driven by real-world examples.

- So it's not a reduction, but on the other hand, I think pluralism has a downside. So for instance, I found an empirically rich account but I lack an analytic account. And my worry is that it tends to make everything comparable, and potentially on an equal footing. So my worry is that it sends out a morality and a pressure to accept whatever

seems natural, right? "This is in accordance with evolutionary or cultural tendencies. And all the technological innovations that emerge."

- The technological innovations are the account that we were most focused on. My worry is that without having an analytic account, just a single rough and ready pluralistic account, anything that emphasizes behavioral modification in particular, is really congruent with this approach.

- After all we know, we know behavioral economics works. Darn good. But just Dunbar's number is a feature of us as a species, I worry that we're losing out on some of themoral claims that might be passed to this.

- And then there's a side issue of a scale. How do we choose what to use at what scales?

- And Bruce provides really nice guidelines. There's a diagram that shows the different societal factors, where they work at time and scale, but it's just a comparative analysis.

- If I were to redo this, what would this book look like if we started from very different conceptual social foundations? I am unashamedly Foucauldian. So if I started from the position of power, and I took all these examples and the richness that Bruce provides, what would that look like?

- So what would it mean? What would this book look like? How would it be different? But if we had a different kind of analytic, I wonder if some cases might help us be able to draw some comparisons. Especially when it comes to questions around morality.

- That's my exposition. More than anything, the book has a clarity that enables us to get to substantive issues, rather than be worried about questions like, what did the author meant here? Thanks Bruce. For a fantastic book and for all your work, really.

Bruce Schneier

- I want to back up a little bit, to give the foundations for where I got to that Quinn talked about.

- My basic starting point is that trust is central for society. That humans as a species are uniquely trusting, that if I was to present this work, a year ago we'd be in the same room, all sitting around a table. And we would be a hundred percent sure that nobody would jump up and attack the person to the next to them.

- And if we were chimpanzees, that would be impossible. We're the only species that trust at the level we do. We trust thousands of times a day, when we eat food, get into taxis, hassle the drivers on the road again and again and again. Society doesn't function without it. And it's largely under the surface, and we don't even think about it. And I think that's really important.

- Trust is actually a tough word. It's an overloaded concept. And there are two basic kinds of trust. There's a personal intimate trust. When I say I trust a friend it's less about their actions and more about them as a person. I trust their intentions.

- But really, often we talk about trust. We talk about the less intimate, less personal trust. I might know them, not know them personally and other motivations, but I trust their actions. I want lunch from a takeout person today? I never met them. But I trust that the food will not be poisoned.

- I might not know you want to steal, but I trust that you won't. So it's kind of confidence and compliance, consistency, predictability. It's a bunch of things that I in the book called cooperative.

- We need to trust people, institutions, and systems. I get in an airplane. It's not like I trust the pilot. I trust the airline and the regulatory system that produced him. I don't trust the taxi driver. I trust whatever system creates him. I don't trust the ATM machine. I trust the entire system and I think that's important.

- So all of these systems require cooperation. Biological, social, sociotechnical. This is where I get to my basic metaphor: in any cooperative system, there's also an alternative parasitic strategy. And that's tapeworms in my digestive tract. It's thieves at a market, spammers on email, people don't pay taxes. And I'm calling them defectors to really invoke the prisoners level. The basic idea is that parasites only survive in the system, if they're not too successful. Like greedy tapeworms, you die, they die. Too many thieves in a market, the market closes down. Too much spam, no one reads email.

- So there's this fundamental tension between cooperating and defecting. And that sort of us as individuals versus us in society. I don't want to pay taxes, but we're all better off if everyone does. And even more importantly, we're each better off if everyone cooperates, but us. I am the most well off if I live in a society where we're no one steals and I get to steal. But if everyone acts that way, societies collapse, and this is where I think we need security to enable trust.

- Security is a tax on the honest. James Madison: if all men were angels, no government would be necessary. And then I go into all mechanisms society uses on itself to reduce, although not eliminate parasites, defectors. Optimal crime-rate is not zero. That's too expensive. But of course, too much crime is too expensive as well. And that's true for credit card fraud. That's true for cheating on school exams. It's true for everything.

- And then I go into all of these four, what I call pressures, that society brings to bear on its members to keep the number of parasites to a small enough amount. A small enough percentage that the system still works. That's what I wanted to add, but Quinn, you want to write a volume two of this I'm in!

Primavera de Filippi

- To me it's super interesting Bruce, because we actually had a lot of discussions about those questions. We had a few papers that were explicitly defining trust as game theory. Whereas you introduce it with game theory but then you explain all the stuff that happens outside of it to actually change the payoff structures.

- It seems that the way you're presenting it is that all the social pressures are tools that enable us as a society to reduce the number of defectors and encourage more cooperation, so that overall we feel comfortable trusting the system with whom we interact. Those are all ways, little tools, little levers, that we can use in order to enable a fruitful ground in which trust can emerge. And those little levers, some of them relate to confidence: the security ones definitely have a confidence-related aspect.

- And so it seems from reading the chapter, that in your eyes, confidence aka security, is something that actually creates the necessary grounds for trust to emerge.

- We had this discussion in terms of, if you focus exclusively on confidence and if you just create a system in which technically you cannot do anything but comply, is it actually leaving room for trust to emerge? Or are actually replacing trust with confidence?

- And in that sense, you just have these highly securitized systems, but that doesn't mean I trust anyone or I trust the system. It just means I have confidence in it. And so the other side of the coin would be that increasing confidence to an extent will actually reduce the opportunity for trust. You're still reducing the risk, but that does not necessarily means creating more trust in the system.

- So can you elaborate on this interplay between the way you see security and confidence on the one hand and trust on the other?

Bruce Schneier

- So I think this is a lot of a nomenclature problem. Trust is a tough word and people who write about all different definitions. To me, the confidence is trust. If a car pulls up in front of my house, a stranger in it, I'm going to get inside, they're going to drive me somewhere. Why do I do that? Because I trust Uber? Because I have confidence in Uber? What's the difference?

- I mean, we could say that confidence is the stuff that Uber forces and the trust is that magical stuff on top of that, but I think we're playing semantic games. I think you really have to shovel all that together.

- I mean, why would I buy a Bitcoin? I have confidence that I won't be cheated. I trust the thing. A lot of it's the same thing, and yet we can talk about different pieces of it. Some of it is blind, some of it, the algorithms, the math, some of it is going to be the governance and all those are gonna work together.

- And that's why I stepped back and used a very broad definition of trust, rather than narrow ones where trust and competence and consistency are different. It's very fluid, and different people mean different things.

- Reading, people who write about trust, they all have different definitions. Some people will say trust, is the thing above all the tech, right? The stuff that you can't quantify or force, but you know, if you're not gonna rob me because you don't want to, and you're not gonna rob me because you can't break the lock, in both cases, I'm going to trust that I can leave my home and it'll be okay.

- And I think burglary is a good example. Most people don't rob my house because it's morally repugnant to steal. Some people don't Rob because what would their friends think? Their reputation. Some people, because they'll go to jail. And at the very top, that's a sliver of people who won't rob my house because they can't pick the door lock. But that door lock doesn't work without everything below it. If all I've got is the door lock, then it's a lot harder.

- And then as we get to more global systems, you start losing morals, losing reputation. All you got is the door locked. In some international systems, because one person does so much damage, we now have to worry about that three Sigma that isn't determined by anything else. And also to your levers, I actually do that in my book. I have the knobs, but it's the exact same thing. And you're right. We, as society tune them. Did I answer?

Primavera de Filipp

- I think that's something we have tried to figure out. The extent to which it is useful to distinguish between trust and confidence. It's useful to understand them as different concepts, but one is having an effect on the other, and I think it's an interesting interplat to analyze, but that's also something that can be looked at from a higher level, by just condensing them into interlocking things.

Balázs Bodó

- I would like to ask a question about the relationship between technological security systems and what you describe as social pressures. I'm re-reading Karl Polanyi's Great Transformation, and what he describes there is that all economic interactions are socially embedded. This is what had regulated the economy before the industrial revolution. This reads very much like your social pressures concept, that practice is socially embedded in relationships in networks, in institutions.

- Now the technical systems that you describe, and also this group is looking at: these planetary scale impersonal technical systems, you also point that they disembed trust from the social relations. The question is what happens then? We know what happens when the economy gets disembedded from social relations, and money and labor and nature becomes commodified.

- And now the question is, can we expect maybe something very similar, when we have the technologies disembedded from trust production and from social relations?

Bruce Schneier

- Well, I think we're seeing that, when you look at eBay's feedback system - its a way to algorithmically make trust emerge. How do I know which sellers are trustworthy? It's feedback. And they spent a lot of years tweaking that to make it work. So I think we're seeing what happens when trust becomes unmoored, and that's really a function of scale. I think we learned through our human history that formal systems emerged when groups got so large that the informal stuff didn't work.

- The reason we have laws is because reputation and things that would work in a group of 50 people, aren't going to work in a city. So you need codification. Even something like criminal codes, which are not laws, but they are rules of a large organization. Code to the underworld is a really good book on how a criminal group establishes trust. Super fascinating, cause they're all criminals. These systems emerge.

- So I think, yes, you want to look at the top end of the scale, and things that have happened in the past 10 years really are interesting in that way. A lot of these decentralized trust systems are a way to, have a lot of texts, so we don't need much of the other stuff.

Eric Alston

- I have two very short points. One is I think, separating out the security vector as creating a set of technological pressures or constraints on the incentives of potential participants is very strong.
- I don't know how well you can see this, but this is a 19th century secured ledger, which is in locks, but not so that you can't see the pages, but so that you cannot remove the pages without the permission of the key holder. And so I've been looking at 19th century patents in a variety of lock and safe areas to understand how those interface with the boundary of organizational forms.
- My suspicion is the effective organizations we have are actually constrained by security technology in important ways. So two thumbs up on that. Small definitional quibble with the way you characterize institutional pressures. You say they occurred during or after the context of a social dilemma or a triggered activity by somebody committing time, or it transgressing a rule of a private organization.
- But as I understand it, deterrence is a really important aspect of codifying any social rules. And so for me, there's definitely a before component that at least is pretty heavy out there in the law and economics literature, as this is why we enact laws, not just to punish, but also to deter others.

Bruce Schneier

- This is the chart that Quinn mentioned. Deterrence is a big thing. A door lock stops you for breaking into the house. Of course, things have a deterrent effect. Maybe not in that chapter, but I do spend a bunch of time on deterrence.
- So is reputation, or anything that involves other people. Or even morals, right? If you think you're going to go to hell, if you do the thing, you will be deterred from doing it.

Eric Alston

- I'm glad you raised morals though, because I think you're talking about a very important class of social dilemmas. But I also think there's a wide swath of social disputes that parties can predict will emerge ex-ante, but they can't classify the nature of the underlying dispute. And I'm talking about incomplete and relational contracting literatures, as speaking to a huge set of incredibly economically significant social disputes that both parties to a transaction understand, that will emerge in the future, but they can't possibly specify ex-ante all of the downstream contingencies associated with their highly complex intertemporal commercial relationship.

- And so by specifying a third party that will credibly and impartially adjudicate the situation, no one could predict would emerge ex-ante, and I would argue that doesn't always have a parasitical alternative. Or to put it differently, doesn't always have a strongly moral implication, if any, at all. So a container ship tips, and a set of containers fall off, and it's an unfortunate incident, but who bears the cost of loss goods? That one? Probably mechanism design. You can plan ex ante. But it's a simplified example, and there's a set of things where we can't predict that both parties are like, we don't want that to happen. In the event that happens, suddenly our interests are in opposition to one another in a problematic way for the nature of our commitment now.
- And so we have to have a class of trust in the institution itself. And ultimately this is characterized as rule of law. So the point I'm making is not original to me by any means, but I do think that it's, it's an important class of disputes that don't always have a parasitic alternative, or strongly moral implications that nonetheless are bundled up in this web of pressures that you outlined.

Bruce Schneier

- I forget if I wrote it out here, but I do talk about adjudicators in other things I write. But yeah, I think that adjudicators are really important. And this is like a trust in a safety net. That if our contract will necessarily be incomplete, some of that will be filled in by goodwill. Some that will be filled in by the judicators that we both agree on beforehand. Even though we do not know the terms of the dispute.

Eric Alston

- To me, this is actually linked to why the rule of law emerged as an institution. When in theory, it would bind against the interests of the most powerful, in very problematic ways in the future, because they can no longer act with impunity, but in a context where no one can identify an impartial third party administrator, that's a context where commitments are very costly because the less powerful, never trust, the more powerful not to subvert outcomes. So the intertemporal nature of commitments is constrained as well as the complexity of commitments is constrained.
- And so there's this interesting way in which I see this as another sort of parallel aspect of trust.
- I hesitate to label it fully a critique. It's more a limitation on the scope of the set of social dilemmas for which there's an unambiguous parasitic alternative, as you've noted. The set you've noted is hugely important, perhaps the most important. So I don't mean to belittle it at all. I just think there's another massive set out here, which is economic commitments that don't have a clear moral parasitic implication.

Judith Donath

- I'd like to push back on the difference between trust and confidence. I think we need to really take seriously the social embeddedness of trust. Trust in that sense is both sort of more interesting, more problematic than it can seem when we just look at it as another way of achieving confidence.

- One of the important things also to keep in mind is the underlying emotional experience of trust versus, confidence that is achieved through technological means or other sorts of security. If I trust you because we have established a relationship that process of doing so, is something that has a very positive affective valence to it.

- There's a whole experiential side to trust that is very pleasant in terms of what it means to be a person within society. If you substitute that with confidence achieved either through some kind of constraint or other form, you do lose out on that side of it. You lose out on a lot of the necessity of the mechanisms of achieving trust that you had had otherwise.

- One way of looking at that is to look at the transformation of something like couch surfing to Airbnb. Where couch surfing involved, investing a large amount of time into developing relationships, making yourself likable, finding other people you liked, in order to trust what had been a stranger to sleep on your sofa or vice versa.

- In Airbnb, you don't have to do any of that because there are constraints imposed from above. On the other hand, that embeddedness of trust has a whole other set of social problems or issues associated with it, or don't want to go into right now. But an obvious one is it's somewhat antithetical to diversity in that a lot of what makes us trust others is the more similar they are to ourselves.

- To the extent, for instance, if you want to have something where people open up their houses to a diverse range of different people. It's not going to be through mechanisms of social trust. It's actually much more achievable through non-emotional technological needs.

- The second bit of vocabulary wanted to push back on a bit was the equivalence of ratings and reputation, because with eBay ratings, Amazon ratings, et cetera, a simple way of looking at that is when you look at reputation as embedded within the society, it's the conversations among members of that society, it's part of the social capital of being within that group.

- The person, if I go to a restaurant and I like it, or I don't like it, and I talk to other people, my saying this place is great or not makes that knowledge part of our circle of friends.

- We'll assume that I'm closer to the people I'm talking to, than the thing that I'm rating. When you deal with ratings online, you have no relationship to the recipients of the information you're being asked to provide for free. So it's a very different experience of solely being about your desire to reward or punish the thing that you are rating. You don't have that same concern with being a trusted radar in a community that you care about.

- And my third quick point is in codes of the underworld. If I remember correctly, it's been a while since I've read it. But I think a lot of his point in looking at communication among a criminal underworld, is that he was really interested in how humans signal. And costly signaling is about communicating in a situation where you don't have to have trust. It can work with trust, but if the signals are sufficiently costly, you don't need trust because the costs embedded in the signals preclude the ability to defect.

- So a lot of the issues in the codes of the underworld, wasn't so much about how people came to trust each other, it was about how they managed to devise ways of communicating that most of us don't have to fall back on because we do have trust, but they were communicating with no trust. And how could they, how could they develop signals that let them do that? So it's a slight point, but it is a different take on trust.
-
- Uh, can I just add a micro piggyback point on this? My feeling at this is that, um, if we look at the incidence of trust and confidence on risk, it seems to me that when we're talking about confidence, confidence exists because you're trying to eliminate risk. Whereas trust actually exists when you're finding ways to cope with an existing situation. So the two can be subsumed because they make you act with whoever you would not have acted otherwise. But the way in which you act is very different. If you act, because you think that is no risk or because you know, there is a risk but you decide to cope with this risk by trusting someone.

- Yeah. And I just want to add to that a lot of what we have to understand about trust is both the cost and the value of developing it. And a great deal of what goes into developing trust is an ongoing series of things, basically risk tests. If you even think about traditions of gift exchange In communities where there's a very strong tradition of that, to some extent every time you give someone a gift and then they owe you something back, you have established this like minor case of indebtedness. That's one way of looking at it, is it's a sort of continual process of risking being ripped off so that you can see the other respond in a way that shows their trustworthiness, and that you keep doing things like that over time, as part of that process of building up trust.

April 8 — Pettit (1995): Cunning of Trust; (2004): Trust, Reliance & Internet

Attendants: ???

Transcript:

Discussant — Chris Wray:

The author draws two distinctions in the field of psychology.

(1) The first distinction, in the 1995 paper, between:
   - on the one hand a person's disposition to fulfil another's express reliance upon them that amounts to a desirable trait, in particular loyalty to another, or virtue in the sense of a moral duty [between 1995 and 2004 he abandons "god-fearing" adherence to "religious norms" for "kindness"], or prudence as, for instance, in the context of a commercial relationship [by 2004 this is framed as perception of the utility of sustaining a long-term cooperative relationship - what I would call frankly transactional behaviour] - these the author places in the category of "trustworthiness"
   - on the other hand, a person's disposition to fulfil another's express reliance upon them that amounts to a neutral or even undesirable trait, in particular, desire for esteem/to be well regarded/to have the good opinion of others/to enhance and maintain one's reputation, whether that desire be "intrinsic or instrumental" - this the author labels merely "trust-responsive".

(2) The second distinction, in the 2004 paper, between:
   - on the one hand, coming to know a person - a "real-world, identity-laden" person - through (a) their embodied expression, (b) their interaction with others whom I "know and credit", and (c) the accumulated evidence over time of their behaviour towards me and others; as this information is available to me through all channels including interaction in person;
   - on the other hand, coming to know a person - a "real-world, identity-laden" person - through (a) their embodied expression, (b) their interaction with others whom I "know and credit", and (c) the accumulated evidence over time of their behaviour towards me and others; as this information is available to me purely by way of digital media transmitted over the internet.

The author brings these distinction to bear on the topic of trust, which he narrows for his purposes - without commenting on trust in any broader sense - to "interactive, trusting reliance":

- interactive: the reliance is expressed or "manifest" to the other

- trusting: the other is expected to attach greater utility to acting in fulfilment of the reliance at least in part because of the reliance, because it serves the purposes of the trusting person.

The author claims that the reasonableness of trusting another for whom the utility of fulfilling that trust is based on a desirable trait is obvious: "to be trustworthy...is to be reliable under trust and to be reliable, in particular, because of possessing a desirable trait";
In contrast, the author thinks that the reasonableness of trusting another for whom the utility of fulfilling that trust is based on a neutral or undesirable trait is not obvious and requires an argument, which the author duly provides:

- 1st premise: an act of trust signals to the trustee, and to witnesses, that the trustor presumes some desirable trait on the part of the trustee such as to increase the utility of fulfilling the trust;
- 2nd premise: the trustee in fact has the neutral or undesirable trait of desiring the good opinion of the trustor, and of witnesses;
- 3rd premise: this undesirable trait increases the utility of validating the presumption of a desirable trait; so

Conclusion: it is reasonable for the trustor to expect the trustee to perceive intrinsic or instrumental utility in fulfilling the trust.

The author then claims explanatory power for the distinction: he claims that trust that presumes a desirable trait that increases the trustee's utility in fulfilling the trust fails to explain why:
- it is a pleasure to be trusted;
- trust is ubiquitous in civil society;
- trust can be created without prior knowledge of the trustee;

In contrast, trust that presumes no desirable trait, but merely the universal neutral or undesirable trait of desiring esteem, as what increases the trustee's utility in fulfilling the trust can explain these things:
- it is a pleasure to be trusted because it is a pleasure to be well regarded;
- trust is ubiquitous because the desire to be well regarded is ubiquitous;
- trust can be created without prior knowledge of the trustee because the desire for esteem is universal.

If I may comment here, it seems that all of the explanatory power derives from the (near-)universality of the desire for esteem, not from its purportedly neutral or undesirable character; altruistic behaviour is to some extent innate in humans; what's new here?
The author's argument for explanatory power would go through even if the desire for esteem were in fact seen as a desirable trait; the argument requires only universality. but sadly the

non-universality of the desire for esteem, or rather the less than universal scope of the community from which a given person desires esteem, means that it will never be entirely reasonable to trust a stranger; this brings us to...

The author's claim for institutional significance of trust-responsiveness:

- Regarding the 3rd premise: the undesirable trait of desire for esteem increases the utility of validating the presumption of a desirable trait - but only if things are arranged in such a way that it is obvious whether or not the trustee does indeed behave in the required manner, without which no esteem will be earned;

But this, I suggest, applies equally to trustworthiness: if it will not be clear even to the trustor whether or not the loyal or morally dutiful or commercially prudent trustee fulfilled the trustor's reliance on them, then there can hardly be said to have been reliance, and thus trust, in the first place. There is a point here to be made about institutional verification, ideally objective and on the public record, of relied-upon action; but I suggest this has greater relevance to organisational learning than to trust

- Regarding the 2nd premise: the trustee in fact has the undesirable trait of desiring the good opinion of the trustor and of witnesses, but only if there is not a division in the community, in particular a division between the trustee and those others, which makes people on one side indifferent to what people on the other think of them

But this applies equally to trustworthiness e.g. once-loyally mutually supportive British families divided over Brexit; e.g. once-kind and morally virtuous Burmese Buddhists now happy to persecute Rohingyan Muslims

- Regarding the 1st premise: three conditions are necessary for the expression of reliance to communicate the presumption that the trustee has some desirable trait that increases the trustee's utility in fulfilling the reliance: i.e. "that there are enough instances of trustworthiness to make it plausible that a trustor should hold such a presumption" (this, obviously, applies equally to trustworthiness); "that the trustor does not have any more salient motives for manifesting reliance" (this applies equally to trustworthiness); "that the trustee is not subject to such pressures to act in the required way that any manifestation of reliance is more plausibly explained as stemming from a recognition of those pressures" (this, again, applies equally to trustworthiness).

Since all the above conditions apply equally to mechanisms of trustworthiness, I am not sure what we learned about institutional governance by drawing the distinction.

Pettit adopts a similar frame in his 2004 paper

Being trustworthy is "being antecedently disposed to respond to certain manifestations of reliance. [Persons] may not be disposed antecedently to display the trait or behaviour you want them to display but they are disposed to do so, other things being equal, should you manifestly rely on them to do so. They are possessed of stable, ground-level dispositions that you are able to engage by acts of manifest reliance."

It is not clear to me how trust-responsiveness, i.e. the trait or disposition of desiring esteem and acting so as to preserve and enhance it, is not a stable, ground-level disposition; it seems to be a near-universal human disposition.
Regarding trust-responsiveness, it's "not that those people are currently disposed to respond appropriately, but rather that they are disposed to form such a disposition under the stimulus provided by your making a relevant overture of trust. You might think that they are meta-disposed...to provide some help you request [etc.]. They may not be currently disposed in such directions but they are disposed to become disposed to respond in those ways, should you make the required overture."
But nor were the "truly" trustworthy persons--the loyal or prudent ones--"disposed to respond appropriately" other than consequently upon the trustor's manifest reliance on them; they were not disposed to fulfil the reliance before they were aware of there being any reliance. Pettit seems to be offering an alternative basis for his distinction: rather than the desirable and neutral/undesirable character of trustworthy traits and trust-responsive trait, respectively, here he argues that the former are dispositions and the latter is a "meta-disposition". This distinction seems empty to me.

The author now brings in another bit of psychology:
"To manifest trusting reliance is to provide normal, esteem-sensitive trustees with an incentive to do the very thing which the trustor is relying on them to do. It is a sort of bootstraps operation...
"The esteem-related way in which trust may materialize depends on its going without saying--its being a matter of routine assumption shared among people--that when a trustor invests trust in a trustee, that is because of taking the trustee to be trustworthy. But isn't it likely that people will recognize that in many cases the trustor invests trust because of taking the trustee to want his or her esteem, or the esteem of witnesses, not because of taking the person to be antecedently trustworthy?
"The assumption is going to remain in place as long as people are subject to the fundamental attribution error or bias, as psychologists call it, and so are likely to expect everyone to conform to that pattern of attribution."

> I'm sorry but it does not follow from the fact that attribution bias is ubiquitous that people are aware of attribution bias or expect others to be aware of it and to hold beliefs in accordance with it. I'm happy to wade into behavioural psychology but I don't see that the author's cherry-picking of one among very many heuristics and biases sheds light on this matter.

The author now brings his distinction to bear on the issue of internet-mediated trust, and introduces his second psychological distinction:
That coming to know a person - a "real-world, identity-laden" person - through

- their embodied expression,
- their interaction with others whom I "know and credit", and

- the accumulated evidence over time of their behaviour towards me and others

is different in the case of on the one hand those ways of knowing mediated digitally over the internet, and on the other hand those ways of knowing including in-person interaction.

> I don't need to have met an academic author in person to respect their work and thus the quality of their judgment; embodied expression is surely valuable in getting to know someone but it seems highly context-dependent - trust for certain purposes may depend more on knowledge of the trustee's embodied expression.

"Second, the evidence available to me as I see the person in interaction with others, enjoying the testimony of their association and support: in particular, see them connecting in this way with others whom I already know and credit."

> By presupposing that it is not possible to "know and credit" anyone through purely online interaction, of course it follows that I cannot gain such confidence from observing a given individual interact with others whom I know - for there are none, by assumption; but surely there is available here the bootstrapping approach of observing over time a number of individuals interact with each other as well as myself, such that my confidence in any given individual increases not only as a result of my direct interaction with them but also their indirect interaction with others in the community in whom I am simultaneously developing confidence?

"Third, the evidence that accumulates in the record that I will normally maintain, however unconsciously, about their behaviour towards me and towards others over time."

> This point seems bizarre. A complete, searchable record of the verbal interactions between individuals in an online community seems far more reliable than my fallible memory of who said what to whom, including conversations that I only partly overheard or missed entirely and know of only through hearsay.

"The striking thing about Internet contact is that it does not allow me to avail myself of such bodies of evidence... these problems stand in the way of my being able to judge that a pure Internet contact is loyal or virtuous or prudent/perceptive

"I think that the possibility of rational, primary trust in the virtual space of the Internet is only of vanishing significance. It is a space in which voices sound out of the dark, echoing to us in a void where it is never clear who is who and what is what. Or at least that is so when we enter the Internet without connection to existing, real-world networks of association and friendship.

I disagree: the same network or graph-theoretic considerations apply in real-world relationships as online: the less connected each of my contacts is to my other contacts, the less secure I feel in my social environment; so just as we see the secular trend of increasing anxiety and loneliness as populations transition from highly connected (typically rural) kin networks to weakly connected networks of mere friends, especially in urban environments where the population density introduces noise that reduces the likelihood that any one of my contacts will

know another of my contacts, so in online forums and social media we can expect one's confidence in any given relationship to be the same function of connectedness of that contact with others whom we also know.

The possibility of artificial personae does little to reduce the evidence of a history of public interaction between others each of whom we also interact with, whether this be in a '90s IRC channel, a naughties web forum, or in the past decade, social media platforms such as Facebook and Twitter. I would go further: the confidence I have in a given individual participating in a densely connected online community, where that individual's reactions and reasoning is tested over time not just directly in dialogue with me but indirectly in exchanges with many others whom I also am able to assess in this way, is greater than my confidence in an individual I have met in real life but who is unknown to any of my other friends. Consistent verbal interaction over time with diverse contacts of mine is an excellent test of underlying character.

"What of secondary trust? ... how can I think that anything I do in manifesting reliance will seem to make esteem available to them, whether it be my own esteem or the esteem of independent witnesses?

"I will have no ground to think that others--other pure Internet contacts---are likely to take an act of manifest reliance on my part as an expression of the belief that they are people of proven or even rationally presumed loyalty or virtue or prudence/perception. I will have no ground for expecting them to take my act of trust as a token of personal esteem.

"If I try to invest trust in others unknown to me outside the Internet, then the profile I assume will have to be that of the idiot or the trickster.

> Obviously the author has had no experience of online communities where decisions need to be made from time to time whom to appoint as moderators or administrators; for sure, there is risk; but owners and superadministrators of online forums are not idiots to delegate authority to those who have established their character through a history of interaction.

Finally, I'd like to comment on the author's two psychological distinctions.

I think the author's distinction in the 1995 paper between desirable traits and the neutral or undesirable trait of desiring esteem is weak and lacks grounding in psychology.

Despite the replication crisis, one of the strongest results in the psychology of personality and individual difference is the stability (and indeed partial heritability) of major dimensions of personality. When the author writes of desiring esteem or to be well regarded, this seems close to the Big Five dimension of agreeableness; look at the aspects of politeness and compassion within agreeableness, or at their finer-grained facets, and we seem to be looking at the same phenomenon.

This does have the necessary quality of universality, though of course individuals vary on this dimension; but it is hard to characterise it as neutral or undesirable: typically, higher

agreeableness correlates with positive outcomes in various aspects of life and work, and is seen as a positive trait. I can't see how this could be negative if the transactional prudence of the counterparty in a long-term commercial relationship cited by the author is supposed to be a positive trait.

The author's attempt in the later paper to distinguish a disposition from a meta-disposition seems empty, without grounding in psychology.

And I think the author's distinction in the 2004 paper between purely online and offline ways of knowing is unhelpful; other factors, such as the density of the network and the quality of the records are more directly relevant; and even on the question of embodied expression, in my personal experience there are times when I have noticed facial reactions from participants in group video calls that I would surely have missed were we all in a room because I simply cannot get such a compact overview of everyone in physical space.

Discussion

Quinn Dupont

- Your second last critique, I read it differently. Pettit talks about "the trust responsiveness that I have in mind is not a trait that many will be proud to acknowledge it themselves. It's a desire for a good opinion of others and so forth. That resonates with me as meaningful psychology. It's inauthentic if I'm just trying to just try to get Chris to agree with me. I'm not really trusting you. I'm just trying to get something from you or whatever. And that's not something I'm proud of. I think that is broadly speaking a negative trait. It might be effective. I may be able to convince all of you that I'm a real trustworthy guy when I'm actually just an asshole, but you know, that's not something you're proud of. And that's the universal claim, at least in my understanding. Are we talking about the same thing there?
- I think so. I'm a bit resistant to the implicit kind of cognitive versus affective, rational versus irrational distinction that he's making throughout. Where either we're going to do something because it's deliberate, very conscious, very intentional on our part, or it's just something that we're doing instinctively or autonomously. He doesn't seem to be very interested in the effect. He's trying to keep this all within the realm of the cognitive. So if you're trusting me and fulfilling that trust is based on my loyalty to you as a friend, that's good. If it's simply being based on my desire to establish myself in the community, then somehow that's lacking merit, lacking virtue. I just think that sort of "it's either cognitive or it's not" distinction isn't really tenable these days.

Quinn Dupont

- I agree with that entirely, that distinction between cognitive or effective seems to break down, but the broader point sounds reasonable.

Chris Wray

- I guess I'm wanting to place: is trust responsiveness more in the affective realm? I'm claiming that this is an aspect of personality. This is a fundamental aspect of personality. In fact, if I'm an agreeable person, that just is how I will behave. So to judge this almost solely on a cognitive basis doesn't seem fair. Personally I'm very agreeable. I will even at times do things that nobody wants to do just because I don't want to let someone down. Maybe that makes me biased, but I didn't really see that I'm lacking in virtue because of that.

Eric Alston

- I'm dissatisfied with the concept of loyalty, because of my difficulty in disentangling it from other motives that I think tend to predominate in most contexts, that people ascribe loyalty to be a motivating factor. This is a specific example of a more general problem I had with some of the analysis, many points of which were made by Chris already.

- But an issue I had with the kind of different underlying motivating forces, that lead one to trust or display reliance in a particular situation was: I thought the analysis was weak in terms of their necessary relationships to one another, whether or not they're compliments or substitutes, or even that level of granularity.

- Think of the bus example. So I'm following the bus, and I want to find my way to the center of the city. In order for me to make sense of a primacy of these forces that lead one to engage a mental model that we're going to broadly call trust. I thought first, if I'm getting onto the bus and there are no passengers, I'm in a very different world than when the bus is packed full of people. I require much less trust when the bus is packed full of people. Because then I think all of the institutional constraints surrounding the bus driver, as well as a lot of social pressures from people on the bus - that to me, aren't even necessarily implicated in terms of professional reliance - make it so that if the bus diverges enough from where it's going, people will start shouting at the bus driver.

- And so for me, this meant there was a primal sorting condition that was implicit in the example that was glossed over. Why am I going down into such granular details? Because I would like a bit more of a hierarchy of, well, this specific trust that I am describing is not necessary in this context, but becomes increasingly likely in this next context, that's almost identical. And this is the economist training that I had. Trying to derive rigorous relationships between explanatory forces is something I deeply appreciate from the economic method. And so if there's only one person on the bus, I would say, I need more confidence that it will be headed to the center of the city, then when it's full.

- Is it just a binary condition that forces that though, as long as there's one person on the bus, do I need as much trust as if the bus is full? I'm not sure. I went through this thought exercise with you to emphasize what my critique was, which was any time, anything approaching logical necessity or sufficiency among the different motive forces for trust, wasn't for me clearly stated. But it may also be that I'm unfamiliar with this particular disciplinary lens.

- And my point about loyalty is, show me a situation where loyalty isn't mixed up with other more primary motivating forces. Where's loyalty to family, absent love? Loyalty to your employer, absent their threat of terminating you in the promise of future rewards? And I often see loyalty as deeply tied up in problematic power relations within organizations. And so it's odd when loyalty is used to restrain those in powerful positions from doing things they want. It's pretty rare that I hear it invoked in those contexts. So for me, loyalty is also tied up in power relations in very problematic ways. But my more general point is: Clarity as to the motive forces for the different forms of trust in any type of primacy or hierarchy among them, would for me have been very beneficial.

Primavera de Filippi

- Just to add on the former discussion I was having with Eric. Chris, you pointed out that a lot of the institutional design that could be implemented in order to favor these meta-predispositions to reliance or to desire the good esteem of others, also applies to trustworthiness. Similar to loyalty, trustworthiness has this thing in which the more I value the virtue of loyalty or trustworthiness, the more I'm likely to also need to be seen by others as someone that possesses this value. So if I give a lot of value to loyalty, it's likely that I'm going to be seen as a more trustworthy person because of my own judgment of my myself, but it's also likely that I'm going to pay much more attention to the fact that other people see and perceive me as a loyal or trustworthy person.

- So while I see the value in distinguishing, because I think a loyal person, or a trustworthy person will always share a lot of importance, um, to whether or not I am being perceived as trustworthy by others. So the more I have this virtue, the more I'm gonna try to show it to others.

- But the distinction becomes interesting where, I think there are also cases in which I might not be giving that much value to trustworthiness or to loyalties, but as rational and self-interested agent, I might be willing to display myself as a trustworthy or loyal person even if its not necessarily a virtue that I cherish on my own. And so I might not judge myself if I were to violate the trust of someone. but I feel that it's better not to violate this trust just because of the vision of others. One is a consequence of the other. So having the virtue, will most likely lead to wanting this esteem of others, with regards to that virtue, but it's not necessarily the case that the opposite is also true.

Wessel Reijers

- I like the paper because I disliked it so much, there was so much I disagreed with. I just wanted to raise a couple of themes for discussion in this regard. When I read the bus example, I felt like this example doesn't make any sense because if you consider this to be a real situation, the response that he suggested it's actually not the response that I think would occur. If you go to the bus driver, and he's already worried because he's being followed by this car all the time. You're saying Hey , I trust you to go to this destination you already wanted to go to, but you'll have to go anyway, because I follow you. I don't think that the response will be of interactive trusting. Because it's a stranger who tells you something strange. So in this case, it would actually just upset the bus driver probably. And then the second thing is that he says that there's a condition where these kinds of trust relations are not the case, when there are certain checks in place, so there needs to be some free room for action. But in the case of the bus driver, I really wonder whether that's the case, because as Eric also said, like, if he deviates from his course because he's a free agent, actually there are checks in place because at some point he can actually be fired. So I didn't understand this example at all. Maybe other people can come and jump into that.

- But then going through the structure of his argument, I think every week I return to this, because time is of the essence. It's again about an argument about time, because the way I understand it, he tries to say that why we trust people is a straightforward thing. And the reason has all to do with time and history, because he says that all these dispositions, loyalty, virtue, prudence, they depend on history. If you have a shared history then, this is a precondition for these things to exist. With unknown people, on the other hand, there is no shared history or is there? Because the shared history that he kind of smuggles in, I think, is this shared history of a gossip network. So what he says basically is that in tightly knit groups there is a shared history because you experienced each other's actions, but in civil society, this is not the case. So you encounter strangers, but then you recreate some kind of shared history through some kind of a gossip network of opinions.

- I think it's an interesting view and we have to link it with other work because in his work on Republicanism, he links to this notion of civility, uh, which for him is a synonym of civic virtue. And he talks about the intangible hand. So that's also why I think he invokes Adam Smith because he wants to be the Adam Smith of political philosophy, I think. He wants to say that just as there is this invisible hand of the market, there's also this invisible hand of reputation. That intangible hand of reputation that creates a gossip network, with which we can enhance our institutions.

- But then this raises lots of questions. What I find most interesting is this notion of the background between desirable and undesirable traits, which he cannot sustain at all. That doesn't make any sense anymore in his theory for many reasons. I think he wants to drop

from Aristotle saying that there's virtue and there's honor. But for Aristotle, virtue has nothing to do with utility, yet for Pettit, everything has to do with utility because at the end of the day, utility is the only true motivator for engaging in the cultivation of a certain trait. Even for these thick traits, like loyalty, virtue, and prudence he says eventually utility is the ultimate benchmark. I'm virtuous because there is a certain utility to be gained from that. And the same counts for a thin notion of trust. Everything becomes instrumental in his view, in the end. So if that is the case, I was really wondering, how do you have a standard for desirability?

- Because indeed as Eric also indicated these thick attitudes, they're often not desirable. The way that he frames loyalty, virtue and prudence, you can have these in a mafia organization and they are totally shit. You don't want those. So in the end, it's almost like you suppose ethics out of the window. There's no ethics anymore. And then that's a paradox because it's just the one thing that he wants to introduce here.

Chris Wray

- It's a paper about psychology and ethics by someone who didn't seem willing to flush out any substantive framework in either of those disciplines.

Primavera de Filippi

- It's funny because he is less focused on a utilitarian and game theoretical approach than the other papers. Yet, it's actually also facing some of the same criticisms of the other discussions, about the fact that indeed no attention is being given to the more phenomenological and intrinsic pleasant feeling of trust. Yet I feel like it's physically much more emphasis on that count.

Chris Wray

- Maybe that is something good to say that at least, even if he doesn't want to admit it, he has at least started talking about things that have a more effective character. And I think that is important. I don't see how we can engage with any issue that has a psychological dimension without taking human psychology seriously.

Victoria Lemieux

- A criticism that I had of the paper or problematic aspect of the paper that we haven't touched on, and I wondered if others had noticed this. He does a kind of a sleight of hand, in toggling between talking about individual trusting relations, such as the individual and the bus driver, for example, and then making prescriptions on the basis of that analysis, to institutions. And I noticed in the first paper, he says there's some extra things to consider when it comes to institutions, but I'm not going to bother with that right now. And then in the final section, his prescriptions, and inferences are all on the

basis of this individual analysis, of individual trusting relations, which he then applies to institutions.

- So I didn't think that was quite right actually. I think you have to really wrestle with the nature of those relations in institutions. That being said, I did think that his foundational proposition that individuals care about being perceived to be reliable and so forth, there's a sense in which it can apply to institutions. At least at the level of the political actors within institutions in the sense that, if I'm a politician, I would care how you perceive me as being a trustworthy individual and reliable, because it would impact upon my ability to be reelected. So I think that there are certain circumstances where you could make the case for what he's arguing institutionally.

- However, I think that I would explain that phenomenon slightly differently in the sense that it has to do with incentives. So it might be that I do not really intrinsically necessarily care, although I might, I might be that kind of virtuous person. But I might be more inclined to open the possibility of explaining it by the fact that the politician wishes to be reelected, and therefore is going to behave in a way that encapsulates the interests of - going back to Hardin - encapsulates the interest of their constituents. So that's a problematic aspect of the paper. But nevertheless, I think there's some room for his arguments, but basically it just goes back to incentives and, I agree with you, Chris, that I think he doesn't make the strong case for this bifurcation that he creates between just normal calculations of trustworthiness and this interactive, special version of trustworthiness he describes. It's the same kind of heuristic. Emotions, and the non-cognitive factors that play into what I consider to be risk calculations that go on very, very fast -- I refer to Kahneman's Thinking Fast and Slow -- about trustworthiness. And it's sometimes based on a lot of information and sometimes it's not based on much information at all. And the sources of that information come from different areas or aspects.

- And then the last point, is a comment on this notion of the file. Cause that's really interesting to me, because my background is archival science. So everything's about the records. So I thought that was interesting, you know, this face frame and file that he refers to. And I do think that there is some way in which having access to a record, some evidence, epistemic foundation, are grounds to assess the trustworthiness of the other party, which, we may have more or less of in different situations. So I think he does kind of pick that apart with respect to the internet. But the challenge there that we face in a complex society is something that the philosopher Gloria Origgi has picked up upon. That knowledge formation, belief formation is a very complex and attenuated process now. And so we can't really necessarily always go on the direct interactions with individuals to form either our own mental record of whether they're trustworthy -- we have to go on the testimony of others. And to have to be able to rely on the testimony of others in a complex society requires somehow that testimony be objectified and fixed.

And that's usually in the form of some kind of record or file. So I found some resonance there with some of the ideas that personally I've formed about how we can form trusting beliefs in a complex society.

Primavera de Filippi

- I'm very curious about what would be the revision of this paper today. I feel like there are a lot of arguments to say that actually a lot of the institutional designs that he is arguing for, actually are over-implemented on the entire infrastructure.

Eric Alston

- A divide that I thought was interesting, that it wasn't deemed sort of definitional or structural is personal and impersonal. Maybe it's just that this nomenclature particularly is endemic in the area I know best, which is economic institutions. But this bears directly on what I find to be a problematic tendency of cabining all economic interactions as something we can carve out as being transactional. There's a way in which they very much are. And I'm very sympathetic to that. My relationship to my brokers of my retirement accounts are really, really granularly defined and incredibly impersonal. In many instances, I haven't met the individuals managing my accounts, let alone the many stages along the way in which these investments are parked yet. Nonetheless, I am interested in these accounts from a virtuous perspective, especially given the arrival of my daughter, my first daughter, November. And so the intertemporal nature of high significance economic commitments has this sort of, I would say there's a nexus to where it feeds into a lot of things we talk about when we talk about virtue: my ability to provide security for my family in the future is I think motivated by a variety of things I would deem virtuous.

- Nonetheless, yes, the strength of these economic institutions is in part giving me a level of certainty today that enables me to be a better fit father and hopefully a better husband. And so I strain a little bit at the separability of our virtuous aims, from our economic aims, at least in heavily market societies, highly capitalized societies, which is obviously the one I'm talking about right now. And so a fair point could be made that a fully socialist society gets rid of this tension or this nexus between economic commitments and our pursuit of our virtues. Nonetheless in market economies that nexus is, I would say pretty important. But I see as a fundamental defining or liminal question is, how well a society's institutions facilitate impersonal commitment credibility. Because I think for our tight knit associational forms below Dunbar's number, a lot of the other things are much more effectively governing interactions. You don't lawyer up and nonetheless, to be able to trust at scale is for me a very different question. It's trusting impersonal commitment credibility among people you may never see again. And so for me, that's a big liminal boundary for describing forms of trust. Or if we're developing a typology of trust, or

we're going to call this reliance, and this true trust; or this primary trust and this secondary trust. And I don't mean to belittle that exercise. I just do think that that liminal boundary between the personal and the impersonal is a critical one.

Chris Wray

- I wanted to go back to Victoria's point and bring up Kahneman's work. Taking that forward to today, now you've got tripartite models of cognition, you wouldn't just distinguish between the sort of autonomous or heuristic, and the effortful and slow, but also between the, the slow deliberate and the whole taxonomy of kind of cognitive failings, with different modes of cognitive failure.
- Now, on whether it's reasonable to trust or not, it's hard not to have to bring in that level of complexity. I don't see how we would expect to, as philosophers weigh in without emotion in that relevant discipline. I just imagine that we'll come up with some simplified conceptual scheme.

Victoria Lemieux

- Yeah I quite agree. I think the heuristics that we use are complex and individualized to a certain extent, and only partially cognitive. They certainly don't reduce to Kahneman's binary formulation.

Eric Alston

- That to me poses the question of whether there is a reliable default set of heuristics. And again, this is me speaking from my priors in a rather self-evident way, which is to say, I think sometimes including Adam Smith gets critiqued for saying self-interest dominates. When I read the argument to be more, a more credible one, which is at the impersonal scale, what should institutions consider to be the default presumption about behavior? And so it is a darker vision of humanity, which is: if we have no information, do we even have a default heuristic? And I think the economic discipline presupposes self-interest as that default heuristic for ordering behavior among impersonal participants.
- I'm not here to argue that that's the one that should always carry the day. We've learned a lot in certain contexts from applying that heuristic. But another way of framing my earlier point about whether there is any type of necessary let alone primary relationship among these motive forces for trust -- I think when you describe the highly individualized, and I completely agree with you and highly context specific, including to time, heuristics that an individual might deploy for us to be able to understand whether they behave with trust or not. But nonetheless, a deep question is, are there any that are reliable defaults? Economics thinks there is one, but that's not a satisfactory answer to many, many people.

Victoria Lemieux

- That's really interesting, Eric. I think we have to think about a couple of things. So the heuristic we individually use when we're assessing the trustee, and then what constraints the trustees behavior -- which I think is more of an institutional question rather than how I personally form my beliefs.
- Now, it's not that my belief formation isn't necessarily conditioned by institutions, because you form expectations on how, in any given scenario, a trustee is likely to behave based on your being part of the same social institution. But I think your expectations are conditioned by institutionally what is considered acceptable behavior. Which is usually defined in law and more informally in norms.

Eric Alston

- I agree. I think your expectations are deeply conditioned by the set of what I call social rules governing a given individual's behavior. But I think you also directly identified one of the most important cuts out there in terms of social rules, which is: is this person subject to a set of laws that I have some vague understanding of, or up to a highly specific understanding of ? Complimentary to that, but even when laws are not present is the other more primal question, which is what set of norms governs this person's behavior? As informed by their virtues, their community, their desire to be regarded well for others, which might well also be a function of their community. All of those, to me, condition behavior where they're like: I would prefer not to break this rule for a variety of reasons that have nothing to do with an enforcement authority that may or may not levy a penalty against me with some probability.
- I'm agreeing with you, this is the type of primal sorting that I didn't feel was very present in the articles we read for today. But I think it is critical to understanding the likely response of a given individual to trust or not, predicated on a certain set of observable external factors.

  Primavera de Filippi

- I think he tackled that, but interestingly, I feel like he used those external pressures as a way that goes against the ability to assess this kind of desire to actually be perceived as trustworthy, or how signaling trust can actually lead to more trustworthiness. I actually like the points that he made, but I feel like it's making too much of a focus on this point of how much the fact that I trust someone will trigger the desire for the person to be trustworthy. And it's not focusing on how this desire to be perceived as trustworthy is related to my own relationship with trustworthiness. You cannot separate my desire of being seen as trustworthy, from my value of trustworthiness. And, and at the same time, you cannot separate it by the type of community of the contextual framework in which I

am. Because if I'm in a community in which no one cares about loyalty, I might care less about being perceived as a loyal person. Whereas if I'm in a community in which loyalty is an important value, I would care a lot. And so this predisposition of actually caring about the esteem of others cannot be assessed just as an individual predisposition, but also needs to account for all the institutional social and potentially economical context in which I operate, which will actually modify, enhance or reduce this predisposition. So when it claims in the last section that external pressures that might lead someone to want to be trustworthy are actually going against the ability to assess these predispositions, I actually think for me, it's inherently connected. The more those pressures exist, the more those predispositions become strong. I actually cannot rely on those pressures, even to enhance this predisposition.

Eric Alston

- I think that's compelling. This may be a useful analogy. For me, loyalty is acting in a way contrary to something else you value because of the value you ascribe to loyalty. I believe that, in a constitutional sense, a bill of rights when drafted, is drastically over-inclusive. Because it is through the conflict of what are initially announced to be fundamental rights, that I think is actually revealed true social preferences - within different specific contexts. And so I'm talking about judicial interpretation of the US constitution, construed very broadly, but I see that active where the fundamental principles come into conflict with one another as actually being uniquely revelatory of the true boundaries of those principles in the first place. And so similarly, this analysis of, is it truly loyalty if it's fear of norm based punishment in your community that is causing you to adhere to that norm that's sufficiently held by other people? To me, if instead, it's your own loyalty that's causing you to behave in a certain way, the norm isn't binding you at all. It's your, it's your own intrinsic valuation for the, for the virtue of loyalty. But if instead it's the norm that's binding you, I would argue it's trading off with loyalty in that theoretical context.

- And so it's this kind of understanding these things as being in conflict is, at least for me, a useful way of probing the boundary between them in ordering our decisions to behave a certain way.

Morshed Mannan

- There's an interesting point here in a footnote where he also talks about not only the perceived trustworthiness of the trustee, but also taking into account whether the trustee themselves feels, a type of shame almost for not acting in a trustworthy manner.I was thinking how it's exactly this sort of construct that is used by microfinance institutions, where rather than using collateral, you rely on shame. You're trusting someone with a

loan and you expect that they're going to feel shame as a way of ensuring that this trustee is acting in a manner that they're supposed to, in eventually repaying their loans.

- Again, it comes back to Primavera's point about how context specific it is, and how this can be deployed in certain situations to achieve certain ends. This was one of the cases where he flagged certain important points and he then sort of abandons it and then it starts to go into another direction. LIike what was said by Chris in his great presentation about how he presents a discussion that is at its core about ethics, but then he sort of leaves us wanting in it. And I don't know if it's the case that we're supposed to read more of his work to get an overarching idea of what he's trying to say, because I noticed that in parts, he refers to some of his other books. So I wonder if we're supposed to see it all together. But one of the main things I felt was that there were a lot of points that I just wish he expanded upon more.
- For instance, the point on the face, the frame and the file was really interesting. And I think that those could be also interesting in trying to discuss this in the context of blockchain governance and to see how we would look at this? To what extent are these three concepts relevant, and would he reconsider them if he were looking at blockchain governance with us?

Primavera de Filippi

- I think it'll be interesting actually to think about how this concept of reliance actually applies to blockchain technologies.

Eric Alston

- I saw a distinction you've drawn between trust and confidence as something that was also problematic in some places of the analysis, which was: it seemed to paint belief and disbelief as a dichotomous possibility in any given circumstance. You either believe or you don't believe. And that was the predicate for the hierarchy of disbelief, which sent me down a very interesting rabbit hole about a logic tree, where if belief - then X, or if not belief - then X, and a hierarchy of those things. But the problem I have is that many circumstances in which trust is relevant to understanding social ordering, aren't predicated on what was at one point determined to be an active form of trust where I am like directly putting my trust in you. I believe the example used was going to the police and filing a claim. It's like a specific affirmative action you take that does require you did or did not trust in a particular instance, and it's conditional on a binary. But it seemed like he was playing a bit of a stutter step between that and other characterizations of social circumstances where a more general class of trust or confidence that a set of things will be the case in a given community or network. It seemed like he was jumping back and forth between those.

- And he seemed like somebody who gave a lot of thought to what he was writing, but I wish there had been a glossary on how he defines some key terms. Because a lot of them are very loaded, or for me are subsets of each other, as I understand it.

- And so a link I saw to sort of the nature of blockchain networks in particular is this intermediate category of confidence, as I've heard it deemed, that is distinct from an affirmative actor, an affirmative act that requires active faith or trust as it's deemed in the article for someone to undertake the act in the first place. You don't file a claim with the police if you affirmatively distrust them. I see that as A very important problem in institutional design for blockchain communities, because things execute with finality in such communities. And so that is interlinked with confidence, which is, we know this will occur because of the way this mechanized almost mechanized system works. And unlike most institutions that we, in the real world, take the time to announce ex-ante; confidence is linked to finality in blockchain networks in a way that inchoate general trust in our institutions is not, because those institutions are default rules that most people want to avoid the application of.

- And so there's actually like a deep, underlying, structural dissimilarity between the application of rule as code - and that gives one confidence. And the credible threat of the application of a default rule that coordinates behavior, which isn't countenanced by the default rule specifically, at all.

Primavera de Filippi

- This is something we've been talking about for the past few sessions. We always end up there. The whole question is, the blockchain is increasing confidence in a particular system, but at the same time, there needs to be some degree of trust in the underlying governance for the confidence to actually exist.

- I'm just wondering if anything that came out of those two papers can contribute to this discussion on trust and confidence in the context of a blockchain. To me it feels that because of the way in which a blockchain is constructed, there's actually very little space for trustworthiness to emerge from the desire to actually be seen as a trustworthy actor - at the underlying governance level. Because it seems that those economic incentives and economic assumptions are pretty much the main motivator. I'm not sure if at least at the mining level, miners care too much to actually be seen as trustworthy. And perhaps, it's more interesting to look at it in the deliberation phase, where you have all different actors that engage into different conversations. Some people might try and promote one possible change rather than another, and whether this concept of reliance actually comes into play or not? Because I cannot think of an example in which I would see this concept applying.

Victoria Lemieux

- I agree with your last point, Primavera that it comes into play more in the deliberation phase. Once the rules of interaction, the consensus mechanism is set, it operates in a kind of an algorithmic fashion. But when those rules are being defined, that you see, some of what he is talking about in this paper -- there's space for that to occur. And I was struck by the fact that he makes a presumption, and he again glosses over it, that there's a cohesive community -- where it doesn't apply that I can rely on the fact that others want to be perceived to be behaving in a trustworthy fashion. Then if the community isn't cohesive then these arguments don't apply. And I think that what you see in the early stages of the formation of an ecosystem is a community cohesion. And that cohesion is based on a shared set of values, norms and principles of operation amongst the social group. And it's a fairly cohesive social group.
- Then as it grows, it becomes more diffuse and you get some challenges occurring. So in the early days, that's where I could see this notion of trust that he brings in, having more influence over how individual social actors form the rules by which the ledger is actually going to take shape.

Morshed Mannan

- I feel that his conception of trust is usefulI in the sense of contrasting with Hardin and Gambetta, and some of the others who we've read before, in their sort of more rational calculation approach defining trust. But while noting this, it's also a bit less useful in understanding trust in blockchain because of the fact that, the rationale that was developed in trying to build a public permissionless blockchain, was more geared towards this rational approach. And so in terms of actually applying these readings, it's useful to look at those who have taken this rational choice approach, rather than this one where it just becomes a conceptualization that doesn't fit as easily.

Eric Alston

- I have a question as to the extent to which the set of heuristics that a person is deploying to decide whether or not to trust, whether this set of participants thinks below a certain level of information. It's a very different category of trust. Because I just put in the chat that I agree, the characterizations of trust here, once a permissionless blockchain network is running, especially with respect to the incentives of network participants, that characterization of this inchoate trust is weird at best, if not outright incorrect. But many users of cryptocurrencies have, I would argue, either no understanding of the underlying technology or even an affirmatively wrong understanding of the underlying technology.
- They have an immense level of trust in that network or reliance in that network, to preserve their store of economic value. It might be herd behavior in terms of just money following money, but in some important sense, if somebody is parking the levels of economic value that the major cryptocurrency networks are currently market capitalized

at, and you posit that at least a subset of those users don't have a clear idea of what's going on. That's an interesting form of faith and trust. Again, this nomenclature is a real problem, but ultimately it's certainly an uninformed belief that their money will be as well watched after, as in other places. I think it's an interesting class of, of trust or faith.

Primavera de Filippi

- I would like to add a little bit of granularity to that, because definitely I think it's faith if we have to name it in some way. But I think it can also be confidence because, in my interpretation of confidence at least, even if I don't know what I don't know, I can still be confident about something. And I think most of those people don't know they don't know. They just expect it to work. And I think the level of trust is not about "do I trust other people to trust that system, so that the value will increase, but I don't think that trust goes into the governance. I don't think anyone even questions the possibility that the system could break, or that some people could manipulate the network in any way. So even if they don't understand it, and therefore they don't have informed confidence. I think you can still call it confidence because they don't even consider themselves to be in a vulnerable position, as in trust. They just assume it's going to work because that's how blockchain works. And they don't need to know exactly how blockchain works. They are confident that it's going to work.

April 22 — Donath (forthcoming): “Is Trust Obsolete?”

Attendants: Wessel Reijers, Beatriz Botero Arcila, Judith Donath, Primavera de Filippi, Paula Berman, Matt Prewitt, Eric Alston, Philemon Poux, Victoria Lemieux, Sankalp Bhatnagar, Charles Nesson

Key concepts:

- Affective trust is the positive feeling you have about another based on the belief that their intentions towards you are good, due to the value they place on their relationship with you and/or their general good morals and intentions.
- The emotional component of trust explains otherwise puzzling observations about our decisions about when to cooperate and our reactions to having our trust betrayed. The sensation of risk-taking or secret-sharing can trigger the sense of bonded togetherness that trust fosters. Neuroscience research is now discovering the neural correlates (e.g. release of oxytocin) of these social tendencies.
- Successful interactions that involve trust are intrinsically rewarding, beyond whatever social or financial benefit the interaction may have yielded: to be trustworthy, and to trust

someone and have that trust honored creates a positive emotional experience, reinforcing that relationship and encouraging future trusting behavior.

- For the psychopath, their lack of social emotions can be adaptive and profitable. Though they are not trustworthy, they are skilled at eliciting trust and otherwise deceiving those around them that they are well-intentioned. The destruction they can wreak is a vivid reminder of the importance of the affective component of genuine trust and trustworthiness.
- Ultimately, it is confidence that makes our interactions possible; trust is one way of achieving confidence, but there are also constraints and sanctions.
- The rise of surveillance and other confidence-ensuring technologies is diminishing our reliance on trust. Hitchhikers talk with their drivers. Passengers in ride-hailing apps, not so much.
- Trust can be insular, sectarian, and antithetical to diversity. The morality of such groups depends on what the surrounding culture is, and why they are operating illicitly within.
- As we evaluate the consequences of replacing trust with other technologies and institutions as a means of gaining confidence, we need to recognize and value trust as an experience itself, and not only a means to an end.
- As our world becomes bigger and our interactions move online, trust alone cannot suffice, but often it can be supplemented, rather than replaced.

Transcript:

Eric Alston

- I definitely enjoyed reading this. I think it makes a very important point. It actually somewhat structurally altered how I view trust in the sense that I think there may be a stronger claim embedded in here, which I'm going to argue for in my discussion. I'm not going to impute it to Judith directly, but I do see an actually very important claim that can be strengthened potentially, and maybe that'll be a point of discussion, but begins by arguing:
- Trust is essential for human development, but the role of trust is changing and possibly declining given technological change, especially as substitutes – that's my term, not Judith's – to trust may be decreasing in cost. She uses the example of credit as well as delivery services in prior periods versus now, and highlights, I would say, very important options to increase confidence, using Judith's term. I would say an interesting question is: to what extent are these other options to increase confidence, complements or substitutes for trust?
- Again, apologies for econ vernacular constantly leaking into my discussions, but in this particular context, I think it's actually quite useful as a means of understanding these other

ways of increasing this critical thing that trust has been traditionally a huge input too, which is confidence.

- And Judith defines confidence as probability of truth or probability that, I would say in my own words, a certain desirable thing will occur, or probability that a certain undesirable thing will not occur. And so in my stylized equation, I wrote probability of truth times one minus P times harm or the negative outcome that the individual is hoping to avoid in a given situation, which trust might play a role and effectively this then pivots after a discussion of how constraints, sanctions and trust are always to increase confidence.
- A comment I had about this was, until you introduce risk later in the chapter, there seems to be an almost perhaps not deliberate, but perhaps deliberate avoidance of the terms certainty. And for me, the language I'm used to considering this type of problem in which arguably removes a role for an effective component of trust at all, which is part of what I like, which is the real meat of this chapter is there's a component of trust, affective trust that isn't captured in this stylized equation of weighing harms and benefits. Unless we can consider that there is an intrinsic benefit to trust that is in addition, that's beyond Harden's definition as you use it. And so effective trust better explains observed patterns of cooperation, as well as the strength of negative response to breach that we perceive when someone's trust has been violated.
- I came away from the chapter unclear whether you're proposing an affective trust as a thicker definition of trust, including all those things that earlier scholars have attributed to trust, plus the affective component. That's one possibility that I came away with, which is if we discuss trust, our discussion is incomplete if we don't treat this important component of trust called affect of trust.
- The potentially more controversial, and ultimately the conclusion your chapter led me to, was that effective trust is the only thing that can be properly understood as trust. And you don't reach that hard. It may be that that's what you're going for. I'm convinced of that fact after reading your chapter. So this may be something useful for us to dig into in the sense that everything else that people have called trust over time, I'm not sure I any longer believe it can rightfully be argued to be trust.
- And in an earlier discussion, I came out somewhat hard against the concept of loyalty and I've since been kind of crafting at least a short note in my head entitled ‘on being disloyal to loyalty’. And I see a very related corollary as a result of this chapter, which is ‘on being distrustful of traditional concepts of trust’. But I want to return to that in the discussion, because you make several more important points about affective trust in particular.
- So whether it be a thicker definition of trust or just trust properly understood is this affective component, it's fundamentally interpersonal as you describe it. An increase in your credit card limit does not bring the oxytocin enhanced benefits of being trusted, of whispering secrets to a trusted friend after a period of lighter conversation that actually precedes the true moments of trust, that yield emotional benefits. And so in order to fully

understand an ongoing shift away from trust to other mechanisms that you list, we need to appreciate the affective component of trust.

- And so this emotional response to trusting and being trusted is linked to oxytocin. And this was your telling secrets example. And so because of our neurochemical wiring, successful interactions involving trust are intrinsically rewarding. Importantly for my understanding, you're saying there's a discrete set of rewards that come to you from engaging in an interpersonal relationship that involves trust that are beyond those calculable in a strict cost benefit analysis. The equation I gave, which is probability of action equals expected benefits minus probability times expected harm of, you know, trusting too much, leaving your house on secured when there are valuables inside, et cetera.
- Another point I think worth unpacking is this component of trust, and I got a feeling for this, it's fundamentally interpersonal. And so a few questions that I might pose is, can this trust ever exist for an organization? And this is where I became a little bit uncomfortable in the distinction between trust and loyalty, because the closest I came up with was, I eventually decided it was more loyalty in the sense of the deep levels of affection that people feel for sports teams. And I think the current discussion in Europe with the possible transition to a super league, but that disenfranchising many teams that people have deep, deep, emotional bonds towards, it at least is an organizational form that people have a deep, I would say some type of emotional connection, but the more I've thought about it, the more I think that that's a form of loyalty as opposed to trust.
- So I'm still on board with your definition, just explaining some of the thoughts that were running through my head, because as you define affective trust, it does seem fundamentally interpersonal, which is, it's that bond between two or perhaps more people. Although I think once you scale into a group of a certain size, the notion of trust does seem weird, at least vis-a-vis the examples you gave. And so because of this neurochemical wiring, successful interactions involving trust are intrinsically rewarding.
- And so I took the sort of main argument of the chapter to be, if you accept the arguments that there's this affective component to trust that yields intrinsic benefits to participants that exceed simply whatever is being exchanged in a strict homo economicus sense occurring between the two parties. Then if you buy that, we need to truly value the role that trust plays in order to understand the coming digital transformation that involves a shift potentially away from trust. And we need to have this effective understanding.
- And so some of the questions I had surrounded: is it a compliment, because in certain situations it does seem like these things enhance one another. And so if I can have enough of an interpersonal connection, fomented in some type of an online community, it may be that that online community is itself strengthened by the ability to include components of affective trust, to the extent that's possible. So I think these questions are very, very germane as we shift into a digital world.

- And so a few questions I finally had. One that might be worth probing is, is affective trust limited to those with whom we have personal bonds? But there's an interesting tension in the definition even there alone because you use, I think, compelling examples of hitchhiking, as ways to discover the benefits of trust with total strangers.
- Two possibilities. You've spent enough time in the car with them to where they're no longer strangers, and they are part of your personal network at that point, and so trust flows between you because they have become sufficiently personable. Another alternative, which I think is more perplexing for your definition, but in my mind closer to the reality of hitchhiking, is trust can occur between people including people you don't personally know. And I think that that's a very interesting aspect of your definition of trust that might be worth probing further. And I reiterate just potentially to foment discussion, is affective trust just the emotional component?
- At one point, you say I'm on board with Harden's definition, but he doesn't include the thing that I'm talking about. But at other points, I do feel a bit like I'm not sure Harden's definition cuts it, at least for me, when it comes to trust at all. I don't buy it. That seems like other motives masquerading his trust, including simple, cold, harsh, plain old economic self-interest. Which is, I don't understand why we would exclude the benefits associated from repeat interactions.
- So I see somebody in a one-shot sense, those benefits are not present, if I think it's highly likely I will never see that individual again. And so to me, in expectation, compared to somebody else that I know I will see over and over and over again, the economic benefits of cheating are much, much higher in the first context than the second.
- And so most definitions of trust to me, collapse to other things, other benefits that the individual is being yielded, except for this emotional component that's intrinsic to the human experience of trusting as we understand it. And so for me, I'm drawing a box around trust and calling it what you described in this chapter period full stop. And so it was useful for me to read because I don't think anything else out there cuts it as trust.
- Finally, I wanted to suggest a potentially different example than hitchhiking converting to Uber riding. I think the right example is couch surfing. Because effectively for me, the problem I have with Uber is it's like a digital transformation of taxi services, but you have a great understanding of the economic incentives that are shaping both the company as well as the driver. And that to me is a pretty stark departure from the understanding of the person offering a ride to a Hitchhiker. The understanding of their motives is in my mind pretty thin. At a minimum you've got, probably wanted to talk to another human being. And that's about all I can say as a baseline motivation for stepping into a vehicle being offered to a hitchhiker.
- Similarly with couch surfing, people around the world who don't know one another offer free lodging in their own home to another person predicated by some type of algorithm on the internet. And so for me, that situation, the person stepping into someone else's home, is a closer analogy to the risks someone is taking when they step into a

Hitchhiker's vehicle. I've spoken for more than 10 minutes. I have plenty more to say on this topic, but hopefully this was a useful summary and a discussion prompt.

Judith Donath

- Thank you. That was great. It's really rewarding to hear when you write something that you've changed how somebody thinks. So it's just really lovely to hear. I do have a couple of responses, so I'll probably go a little bit backwards in time.
- In terms of couch surfing, I agree with you. I have a hugely long discussion of that. And part of it was, this is a short version of what I've been writing about trust. And my couch surfing example got so long and it talks about, um, it starts with Patrick Lee fervor's story of walking across Europe and staying in houses and you know, how people set up trust and the hitchhiking one is structurally pretty similar, just a little shorter and more concise. But I agree with you that I will probably use in the book the longer example. But there's a lot to get into with what happened to couchsurfing.org, like how it went from being ideal to not being functional, which is actually really interesting, but I just thought it was too long for this.
- So mostly I think we agree more than you recognize, which may be a fault just in my writing, but what I think is also not quite clear is, I see two basic ways in which you can think of who you trust. There's a trust that comes from truly personal bonds. When you get to know someone over time and you trust that specific individual. There's also something that is probably a very similar kind of trust. But this is also where it becomes very problematic. That comes from our common membership in different groups, which is why I also wrote that trust in some ways is inherently anti diversity, in that a lot of what we look for particularly with strangers when we trust them is to what extent are they similar to ourselves?
- Looking at Gambetta's work on taxi drivers, what they're basically looking for is somebody like themselves, like in Northern Ireland, the Protestant taxi drivers try to avoid picking up Catholic fares and vice versa. That was like a pretty literal, you know, undeclared war, but a lot of our sense of who is trustworthy in situations comes from that.
- So to get to the sports team piece. I just want to break down your question about organizations and also into two pieces, and the teams are good ones for that. There's interesting research about trust again, where people will trust someone who's a fan of the same team, just because they're a fan. So if you're here in Boston, there's people who will trust someone, who's a fan of the Patriots, more than someone who's a fan of the Bills. Nothing makes that person more trustworthy. It's just that sense of familiarity. So things like teams or common memberships, or all kinds of little things that people have in common in all different ways, go into people's assessment of who is trustworthy consciously or not.

- In terms of whether we trust organizations or non others, that's also something that we write about. I think it's personal, but to be personal, it doesn't necessarily have to be a person, but I think it's why a lot of companies that want to be trusted end up investing quite a bit into the creation of some kind of mascot or personified thing. Geico wants you to really identify with their little lizard, there's Ronald McDonald and smokey the bear and all kinds of things like that. This intuitive understanding of this affective version of trust is what makes a corporation realize you're not going to trust a nameless faceless thing, but to the extent that you have a human representative or a cute animal representative that can be trusted, it will set off those similar feelings of trust. Just to understand, in the book the chapter that comes after this is about the manipulation of our trusting by artificial entities, like AI systems. So it's sort of the flip side of it. I would definitely say that you can trust things that are not human, but it's mainly because our mental representation of them is as a relatable other being.
- Then that sort of distinction of, what is it when you have those repeated interactions, but there isn't trust? And that's why, um, I was interested in, in bringing up the examples of the con artists and psychopaths, just to show that you, that the ways in which trust, um, assists human interaction really is predicated on this underlying chemical reaction that in it's, we can see examples in its absence that you can perform repeated interactions, but have none of that, you know, you're calling it loyalty, but that sense of empathy that leads people to be trustworthy and trusting that you can act repeatedly. And that's an example, you know, that we can see of people acting in regular situations, but without that effective part of trust. So I'll open it up to others here.

Eric Alston

- I have one last discussion prerogative question, and this surrounds in-group out-group concerns with trust in particular, but I kind of see a bootstrapping hierarchy of expectations about another's behavior.
- And the easiest one to suss out surrounds homo economicus, which is, does this person's economic interests sufficiently correspond with what I want to happen in this situation? And in many instances, I would argue that that is sufficiently transparent and clear, that no level of trust is necessary in the sense that I know you're going to benefit, I know I'm going to benefit. We don't need trust. And so it's, it's strange to say, well, I trust this guy's economic incentives that he's going to do the right thing. That's like, you're saying, I see a central to your point is trust is so much more than just faith in somebody's underlying incentives to do the thing you want.
- But in group, I see the problematic nature of homophily as resulting from a deep evolutionary response surrounding uncertainty as to shared values. And so if this person is clearly a member of my group, the likelihood that they share the same set of underlying values is much higher, at least in expectation than somebody I've never seen before, that

doesn't speak the language, and is dressed completely differently than somebody in my in-group. This is a problem I have. And I'm finally getting to my question.

- If two people have very closely shared values that would dictate that they would behave in a trust appearing way in that particular circumstance, is that trust though? To me, if they're intrinsically motivated based on their own value set, to engage in reciprocal cooperative behavior, that doesn't seem like they're motivated by trust. Because as I understand, you're, you're, you're saying trust comes from or to put it in. It's easier for me to put in loyalty terms, which is I see the fans that are loyal to the teams that lose constantly, that's true loyalty. There's the term of fair weather fans, which is, Oh yeah, they're winning. So you love them, but you wouldn't stick with them if they were losing. So you're not actually loyal. In a similar sense between two people, if their values get the trust looking outcome, is that actually trust, or is that them behaving voluntarily under an incredibly similar set of values that just leads them to appear trustworthy?
- But a possibility is from your arguments that yes, it's often correspondence of value, but trust yields additional benefits. And so it is. I, I guess, where I struggled in this made me want to read more of the books. So I guess it's a good chapter, it surrounds that question, which is, is it, is it values are precedent to trust or is it the case that shared values are somehow a component of the expression of trust itself?

Judith Donath

- Um, well I think part of the difficulty here is I think trust is a component of your relationship with another being, we'll say a person for now just for ease, but it could be more general than that. So it's a component of your relationship with them. So if you're in a, if it's someone you really trust and in your ordinary course of life, you may spend a lot of time with them in lots of situations where you never actually have to test that trust. So sort of like what you're saying, you don't actually need to trust you, you do the same thing. You do all this stuff together. Everything's smooth. If you don't trust each other to be the same, you have the same values.
- The trust comes into play, or you're testing that trust in those situations where there's a possibility that defecting would be profitable. And you're trusting that that person won't. So it might be that you spend a lot of time in an easy situation, but then some rivalry or competition comes up and how will they still, and I think you're right, that loyalty is very close. We are tied to this, how do they behave? So I think that's why a lot of writing about trust deals a lot with that question of risk, and where things like gift giving and these other sort of artificial situations that cultures end up setting up in a way to test your trust.
- So say you're in an important situation where you have to see how trustworthy someone is, but things like, you know, if I give you a gift, does it get reciprocated? Do you reciprocate favors? A lot of these cultural ways of putting yourself in some vulnerable position to someone, even though it's sort of a toy example, I think are ways of saying,

let's have a situation where there is some risk here so that you have that opportunity to show that you are trustworthy. It continues to build that trust, which then might be in use only in some exceptional circumstance. Does that answer your question?

Eric Alston

- Yeah. And it seems to have an important implication that I was about to throw into the chat, which is that the true exercise of trust must then be costly for at least one of the parties. Not on net, because your argument is the affective component brings sufficient benefits to overcome the forgone benefits that they would get from breaching. But it does seem to follow from your definition that for at least one party, trust would be costly on the net, were it not for the affect of benefits that they receive from engaging in the interpersonal relationship components of trust.

Judith Donath

- Right, so you can have trust without ever happening to be in a situation where that does come into play, but it would explain the times when people behave that way, that it makes sense to them. It's not an irrational behavior. It's rational given that component and definition of trust.

Eric Alston

- Although I would note, we seem to have gone back to homo economicus.

Judith Donath

- Well, I would say it's a post homo economicus definition of rationality. I think basically it's saying that in fact, all our behaviors in everything we do really comes down to what our subjective experience of the world is. And looking at that subjective experience through how it plays out in our minds emotionally is actually what we experience. And it is actually that's actually the costs and benefits that we have in the end. Money is only a benefit to you to the extent that you like it. And if you don't want it, it's not valuable to you. It doesn't matter if it's a dollar or a million dollars. So I think the problem with homo economicus is more that the behavior isn't based on external measures of cost and benefits, it's based on the subjective experience of it. And by that definition, you can say, everyone is rational, no matter how crazy seeming they are, what you need to understand is what are they experiencing and expecting that makes them see that path of behavior as the rational thing to do?

Primavera de Filippi

- So I just want to add one thing, because I'm not sure I fully agree with the idea that trust necessarily entails trust on the trustee. That would be the case if trust only and exclusively required the trustee to act in a way that she wouldn’t have acted otherwise. And therefore

it's costly because it's constraining her action but the benefit is higher. I would disagree with that, and I would like to think that you can trust someone even when you know or you think that this person will anyhow act in that way. And that's pretty much the most obvious way to trust people. Because I know this person shares specific values, and even if it's in their own self-interest, to me this actually reinforces my trust in this person because there is no cost for me to put trust in their hands. And because I know we're aligned and have encapsulated interests. So I would not want to remove that category of interpersonal relationship from being a trusted relationship, only because it doesn't necessarily entail any type of trust on the trustee.

Judith Donath

- I think what we were trying to say is that trust is the component of the relationship. And then there's different situations that the trustee and the trustee can be in. So you can go along and be in lots of situations where nothing is costly in some ways. The situation in which there is a potential cost that only we're behaving in what we would consider a trustworthy way makes sense only if you count that affective side are situations that prove the existence of it, but the absence of them doesn't mean there's an absence of trust. It's, you know, the trust is in the relationship, not in the situation, but the situation is what lets you see if it's real or not.

Primavera de Filippi

- Right, I guess it's kind of weird and paradoxical in the sense that, the more you know that your trust will be a weight. Like if I entrust you with something, and I know this is going to be a weight for you. It's this paradoxical outcome, because on one hand, it proves that I really trust you. Because I know that even though it's slightly against your interest, you're still going to act trustworthily, but at the same time, it also might potentially reduce my trust, because I know you might actually have an incentive to breach it. So it's a disclosure of how much I trust you, but it also might reduce my trust in you. But I trust you so much that even with this reduced degree of trust, I decide to entrust you with something.

Judith Donath

- Exactly. What you're talking about is putting yourself in some state of vulnerability to another. And whether that turns out to be okay, because they actually are trustworthy and then that can strengthen your trust. Or you find out they weren't trustworthy and then you trust them less, but the risk or that vulnerability is a way of effectively sorting trust. And particularly in reciprocal relationships, going back and forth and repeatedly having been made vulnerable and having it pay off in the end are ways that strengthen trust.

Primavera de Filippi

- Yeah. So essentially if you are aligned with me, my expression of trust will not necessarily change, or increase or decrease the amount of trust I have for you. I would just keep disclosing that amount of trust. Whereas when I disclose trust and entrust you and this entails that you might have to constrain your behavior in order to act trustworthily, then, depending on whether you do act in compliance with this trust, you might either increase the amount of trust that I have for you or potentially decrease it. So like it's a more interesting exercise to observe when trust is given to someone that has not aligned interests, because there's feedback, either reinforcing or undermining loop, as opposed to when basically I know that I can trust you because you, you anyhow, will do exactly what I trust you to do. And therefore there is no feedback loop because I cannot verify whether you can act trustworthily because of me or just because of your own interests.

Judith Donath

- And that is something that people do in everyday life. Like if you, if you had an assistant, you might, you don't know them that well, you don't know to what extent they're going to be trustworthy. You could let them do things. So if you have a child, you're not sure you trust them to go to the corner by themselves. You might say today, you get to go to the corner with a dollar to buy something, you know, and in a week you can go three blocks away and you know, the next time you can go 10 blocks away. I think there are a lot of times where there are relationships where people extend things like that. Or, you know, you, you have a certain level of trust that you build up over time.
- And that's sort of one of the issues with the difference between that and the repeated interactions and both the example of the con artists and the example that, harden used from the brother's karamazov, of the repeated interactions with the kind of psychopathic, um, man who said, yeah, we'll keep doing this. I will keep doing this. I will keep doing this. And then when I decide to stop our repeated interactions, I will do the economically rational thing and keep your money. Because that's the most profitable thing to do. But a person operating in a realm of trust where the relationship also maintained a measure of importance to them would not have behaved that way.

Wessel Meijers

- Thanks Judith for the great chapter, it was really fun to read. And Eric, thanks for the great comments. I have two points, basically. One point is about the examples used because they are great. And, um, I, I can't help just to talk about my own hitchhiking experience because it's just great. What I was wondering is, even though you give it as an example of the similarity between the two friends kind of whispering the secret and the choice of hitchhiking with a stranger to my mind, actually, they're very different.
- So I'm not sure what is the difference in kind or a difference in, in, in degree to some in some way, because this notion of similarity is stretched quite, quite far, I think, because in

my experience with hitchhiking, for example, like the people with whom I was in the car, like similarity is not the word that described that. The people who didn't speak my language, uh, or, or who were from such different backgrounds, political persuasions, um, in so many different ways are these people different, that similarity does not really cut it, you know? And so I was wondering whether there's actually not, um, two different modes of trust at play here, because when I think about, for example, hitchhiking. Um, trust is immense. Um, and, and it's more of an exclusive exercise. So for example, when we had some of these kind of strange rules, for example, you know, like you would not, you know, you would not ask somebody on a, in a, at a gas station, you would not ask somebody, uh, who has a white van, for example, as a rule, but sometimes you would, but it's kind of like setting some boundaries for a huge space of trust, because I guess once you go into the car with someone, you really pretending, as if this person is benevolent, more in the sense of like what we discussed last time, I think, and when we discussed this paper of Pettit, when he has his, but there's this sense of benevolence that is kind of this ubiquitous feature of society.

- I went one time, for example, I was in Barcelona. I was hitching with a drug dealer. Not really intentional, but it happened to be like that. And the guy was like speeding on the highway, it was super crazy. And he was like, handing out drugs while we were, he was like putting us through the city and it was making some, it was crazy. I was thinking like, does he have a gun or something, but, um, but even in that case, you can have that there remains some trust that is enough to, for you to feel like, you know, I didn't feel at any point like, Oh, we were really in danger for some reason, it's really strange. Um, so I thought maybe that these two examples, the two friends and the hitchhiking, actually do point that maybe on the one hand, uh, this sense of friendship where there's like arrhythmic equality. On the other hand, maybe what you could say, like a sense of justice, where it's about, um, you know, you enter into an equal situation per se, because there is asymmetry between you and hitchhiking and the person in the car, but still through the sense of justice, you can create this kind of common ground of trust that helps you through the sort of drive.
- The other point was about emotion and the way that emotion is captured. So I'm very reluctant to accept any kind of idea of emotion that is based on biochemical processes. Um, because the sort of causality here is questionable. I think, whereas it is wonderful that we have as humans, kind of these biochemical processes in our heads to give us like biochemical rewards and punishment and say, um, but the question I think here is, you know, what, what, what is the causality involved in a situation like the two friends telling each other, the secret? Like is the emotion really caused by this hormone that releases, or is the hormone that releases an effect of the interaction? Uh, so where does emotion enter into the picture basically?
- And also had a question of like, I mean, of course there are interesting outliers, but if we take this kind of dirty monistic vision I think of emotion as a biochemical process, it does

stand a bit in the way of choice and the way that choice affects emotions. So I'm just thinking, for example, like, uh, like buddhist monks, for example, uh, through meditation and through practice, um, affect their emotions in a sense. There's like a different kind of causal chain there. Also in this sense, we are able to affect these kinds of biochemical, um, processes that go on in our minds. So I would just question: What is emotion really? And what is the causal relation between these biochemical processes and emotion as such and trust?

Judith Donath

- Okay. Well, that's, that's quite a lot. So I'll just try to take a couple pieces. I think it's useful to look at the underlying biochemistry without saying it's totally deterministic, but I do think at the end of the day, our experience of emotions is caused by and causes different biochemical reactions in our brain. I don't think knowing the underlying anatomical or chemical or electrical or hormonal reasons why particular things feel a particular way makes it any more or less deterministic. Existence of determinism and freewill is a little bit separate from that, but it is a reflection of what is happening. I think it's also useful because like the examples of the psychopaths, it can show that there is a distinct thing happening in certain cases or not. The question of to what extent does the chemical reaction cause, or is caused by the interaction, the answer is yes, it's both.
- We can get into at length sort of how much your expectation of something happening changes, you know, with the emotion that you have. You clearly like hitchhiking more than I did. Because I found it terrifying, um, but there may be differences in situation and person. I would be afraid as soon as someone said, Hey, we're going to hitch home from the beach. I would be afraid before we got into the car and be afraid before we got on the highway. So some of the emotion is anticipatory. You know, if we got picked up by the drug dealer, I was not having the good time you are having. I think the fact that it's based on biochemistry is an underlying understanding, but whether we understand that or not doesn't change its existence.
- You were asking at the beginning about the difference between, uh, in terms of similarity. I would argue that similar doesn't have to mean you're similar in language or you're similar in other, in, in sort of obvious ways. We're living in a particular period in history where, as our world becomes much more global, right now we're struggling with our tendency towards really looking at homogeneity as the way we form groups. And we happen to be living in a historic time where that is seen as a serious problem. Historically, as humans that has not been pointed out as a problem, it's been considered normal. I think the reasons for it being something we're trying to overcome in this historic moment are very good, but it still means we have a considerable amount of human history and nature to deal with. And particularly in terms of how we form our ideas of who we trust.

- Your example about feeling trust in a place with someone who's very different than you who's putting you at high risk is a situation that not everybody would find trusting, but I think it also gets back to a lot of the interesting accounts, both of hitchhiking and couch surfing of people who found those spaces and experiences extremely highly rewarding experiences of trust, because they were such counterexamples to that notion that it was about trusting people you were exactly like. I think a lot of it was about this real fascination and rush that comes from finding yourself in a situation of trustingness with somebody very different than you. And when I started reading accounts of people writing about couch surfing, that came up over and over and over that what they were looking for was this experience of meeting someone who seemed really different, putting themselves in a very vulnerable position by having them stay in their house or staying at their house, and then in the course of their time together, discovering how much they liked each other. And it was that sort of creation of trust in an unexpected way, in a sort of relatively risky situation that people found so compelling about it. Does that answer your question?

Primavera de Filippi

- The couchsurfing example is interesting. I agree with Wessel at the meta-level because the first act of actually joining or accepting a particular couchsurfing experience, to me it's not about expressing trust in that person. I don't know who that person is, I have no reason to trust the person. And so either it's because I trust the platform like Airbnb. Or it's that I don't really trust that person yet, but I trust the benevolence of humanity that probably this person and I share a particular set of values, but the satisfaction is different from when I share secrets with my friends, which is I'm expressing a particular type of trust and additional vulnerability.
- In couchsurfing, the satisfaction comes from zero to whatever, I'm bootstrapping trust and that's what's really exciting. If I just increase trust a little bit, it's great. If I start from nothing and add a little bit of trust after the interaction, it gives more oxytocin I guess.
- But it's interesting to distinguish that the initial step of couchsurfing is less of a specific interpersonal relationship of trust, and more of a localized form of trust for people on the couchsurfing platform, or people who are willing to pick up hitchhikers, and then you create the trust relationship.

Judith Donath

- Exactly. In that case you're not looking for someone who is on the same team, but in a way couchsurfing is your team. So that's the initial trust based on similar group membership, and in this case it's people who are on couchsurfing. It's enough to trust them more than you trust a random person walking down the street, and then from there, you can actually develop personal trust with them, from talking to them and seeing their house and whatever.

- I realize I didn't address one of, or I think it's underlying what you're saying in the whole notion of bootstrapping, it was one of the points Eric made and I forgot to address it, was how do I see constraints and sanctions and trust working together. And I think that's one of the really interesting ones in general, because I think that's where you get into what, what can you do as designers or futurists or critics? And I think, yes, the point is that these are all things that together go into giving us that confidence. And in sometimes explicit ways you can trade them off, you can have more constraints or sanctions and then you need less trust. To the extent that you want to have less sanctioning, you need more trust.
- And so part of what I'm arguing here is that as we move into a world that I see, because of the increasing ubiquity of surveillance, is making it increasingly easy to get most of our confidence through external controls particularly through sanctioning, there's an argument made that trust, complicated, difficult, problematic as it is, has value in itself. And that we want to create worlds and situations where there is that space for trust, that we don't want to just foist all of the confidence-obtaining into the world of sanctions and surveillance, even though that is possible for us to do today. At the extreme, the Airbnb and Uber version is that world of sanctioning where there is constant surveillance on all the parties. So that nobody's going to misbehave in any way, because they're on camera constantly and they are being surveilled all the time. I'm not saying that's exactly what happens now, but you can move into that world. And what I'm saying is that there's actually a high value to trust and to maintain that.
- But one of the things we can think about is that notion of trust being something you build, starting with small risks and moving on to bigger and bigger risks. To the extent that we want to articulate that, we may want to build systems where you use that ability to start off with fairly low risk situations and then start getting rid of other constraints and sanctioning, and say that building up trust is a goal in the systems we want to build, as opposed to something that we want to just eliminate.

Primavera de Filippi

- What you're saying actually confirms what Eric said in the beginning. The question of, are constraints and sanctions a compliment or a substitute to trust? If the answer to Eric is that they are complementary, it also goes against the statement that I only trust people that have a cost the trust they give. So to me, if there is a system that is highly constraining, with a high degree of constraints or sanctions, but if you are a person that I actually trust with my life, I still trust you with my life. And so I know you are constrained, and the fact that I'm expressing trust towards you is not removing the degree of trust, but it might not contribute to enhancing it. Right?

Judith Donath

- Yeah. That's what we were saying before that the trust is in your relationship, but it doesn't mean that in all your dealings with that person, you're testing the limits of that trust, you often may not be testing it at all. You may never need to use that trust, even if that is that component of it. So if you trust someone and they're highly constrained, you can still trust them, but in conditions and that sort of constraint, you can't build trust in it. You're not strengthening it. And this depends on the situations and everything, but it does seem that having ways of continuously reestablishing trust is what people do in a lot of situations. So that if you put yourself in a situation where you never need to actually fall back on that trust, it does kind of wither away.

Beatriz BOTERO ARCILA

- Thank you. I think this is amazing and I was so happy to read it, but the conversation that you just had sort of changed my question, and made me rethink it, but I'll still throw it out and see. So I was very interested by the turn to surveillance, and I'm interested as well in where do you locate power in the mechanisms that allow for exchange among strangers. So first perhaps in more primitive terms, but that's not the right word, it's maybe kinship or the fact that you look like me, that you're one of mine. And, and that sort of allows me to assume that the risk that still exists in any human exchange will not take place. And then maybe in surveillance, that's exactly what it is doing, right? Like it's creating an alternative mechanism, but there is still power in those relationships. So the platform became the broker of the possible exchanges and the same is true maybe in kinship. I'm curious. Maybe, maybe it's not even a question anymore because you were just talking about it, but I don't know how you see that interplay that transformation in the power of relationships.

Judith Donath

- A writer I like is Richard Sosis, and he writes a lot about religion and signaling. So a lot of this comes out of work I've done on signaling, which is effectively a constraint, of signals that are so costly you can’t defect. But if you look at really small close-knit high trust communities that Sosis' research tends to deal a lot with close-knit Orthodox Jewish communities, which are very, very high trust. And there's this sort of constant demonstration of your trustworthiness through engaging in all kinds of very costly rituals that you would only be doing if you were really committed to being part of that community. And within it, within the circle of that community, people are very, very trusting of each other. And a lot of what he's interested in is for instance, like the Orthodox Jewish communities that are in the diamond dealing business, where they will just say, like, here's a million dollars of diamonds, you know, catch, take it for the weekend. And there are such high levels of trust within that community, they can operate very cheaply with each other because there's very high levels of trust. That's all very well and good.

- If somebody defects and proves to be untrustworthy, then you will see that the community also functions as a surveillance state with very heavy sanctions on them. So it's not the simple sum of saying, well, there's no consequences to breach of trust. In those cases, there's still this backup. There's still going to be a lot of sanctions. And part of the issue with sanctions is that in general people would prefer that the bad thing not happen. So a lot of the power of the power of surveillance is in prevention. But people know that it's there. And again, I think a lot of what's important in terms of thinking about how we want to live as human beings is that some of these things, if you don't deal with the affective experience of living in that place, um, it's, it's hard to make sense of them because externally they may look the same.
- And it's another reason the Orthodox Jewish communities work for me is a good example because you have, you know, maybe the 80% of the people who are just all in on this, like they're part of it. They're actually real believers. It's not particularly painful for them to be part of that community. And then, there's a set of people who live in this community who feel the constraints of it really heavily. Like they're just, for whatever reason, they're not comfortable. They've come to believe they're sort of stuck in this community and it's pretty painful. You can have the sense, you know, you can imagine that there's two people in this community who are behaving in the exact same way, they're both doing the right thing. But one of them is having a pretty good time because a lot of their motivation is in this space of trust. They're being motivated by internal emotions that are fairly positive, which is one of the things that's nice about trust. It gets you to behave in a prosocial manner by making you happy. And it makes the other person happy. And the whole experience is pretty good for everyone. You can get a lot of those same behaviors by sanctioning people and saying we're going to torment you, if you don't do all these things, you're living under threat, you do the exact same thing, but you're doing that same behavior out of fear of reprisal. So while the behavior is the same, your experience of doing it is much less fun. In the end most of us would rather be in the first camp.

Philemon Poux

- First of all, thanks a lot for sharing this chapter because I find it super interesting to read about affective trust, and I was really convinced by the definition you give, but I just want to react to one part of your chapter and then add my own hitchhiking experience.
- I'm very convinced by the the fact that there's more to trust than just a cold blood calculation and that the oxytocin or feeling good is happening when you trust someone, when there's a trust relationship, but I might not be as convinced by the argument, uh, that there's, uh, decreasing trust, um, in the society. You cite Putnam and he's been very foundational in the saying that social capital has been decreasing, uh, in Western world. And I've read a very interesting article called 'the theory that won't die: from mass society to the decline of social capital' (Thomson 2005). And basically it shows that there has

been a continuum in saying that there was a decline in trust and social capital, uh, generally speaking in the US, but they could not find any trace of evidence that was convincing for that point. And I believe the point that you make actually is, um, going in that direction. Basically, if people feel something more than just interest from trust and derive pleasure from it, they will look for trust somewhere else. So my point is, I agree with you that, um, AI and, uh, ITs will make a lot more, um, areas less dependent on trust. But then won't we turn to closer areas to still feel the need to rely on trust, because that's something we need, we need to be happy basically.

- My other point was the transition from hitchhiking to Ubers or renting or paying for your ride. I visited, uh, Romania in 2000, 2004. So at the time it was transitioning to where it's a more capitalistic economy and, but it wasn't reaching the levels of the developments that, um, we know today. And then at the time, very few people had cars and there was no network of public transportation. So the main mode of transportation for the Romanians was hitchhiking, but they didn't really have a choice, so they needed to hitchhike and where's, they could, they actually paid what they could pay, um, for the ride, but still couldn't, um, decide which car to get on, to get onto via an app or to rely on an institutional designs, public transportations. And I think it's pretty interesting. It's this transition system where you have generalized, uh, hitchhiking, but it's not trust anymore because you don't really have a choice to do so. And I don't know where I'm going with that, but I just sort of did that. It was, uh, interesting for example.

Judith Donath

- A couple of quick, quick things. One, most of the people who've looked at the decline of hitchhiking, one of the big reasons they give in the States is that cars just became more common. Um, and so then the people who didn't have cars were seen as increasingly outside of some norm, but, um, you can still have trust if it's forced. You said they have still to decide which car to get into. There's still some possibility. Like if you feel like I have to get to work, I need to get in this car. And the man driving this car is like wild eyed and he's holding a knife and, you know, he's got someone held hostage in the seat next to him, tied up in ropes. I'm just not gonna go to work. You know? So if you have to, you still have to use trust, it's just where you draw that line that will be different.
- And personally, I don't like Putnam's work very much, it was more that I was trying to say that there's certainly been a lot of reports of it. And Putnam's is the most clear, though I think his tends to be a book about people who have stopped going to the sorts of organizations I'm used to them doing, and so it doesn't exist anymore, and not being open enough to say it's shifted into different places.
- Barry Wellman is a sociologist who writes a lot about changing into networked trust as opposed to older forms of trust. So I think it's true that people do that, but I also think we can see measurable changes in how much people need, like even like you're hitchhiking,

but the difference between having to trust people versus having it be an optional activity that you choose to do.

- If you've read Brave New World, there's this interesting notion where every week, you know, you would get this dose of extreme emotion because they had created a world in which they're just, everything was just nice and even. And there weren't any extremes of emotion, but people really need that. So like once a week they would just be injected with it to have that experience of extremity. You know, you could say in a world where people are dealing with all kinds of terrors and everything, they may not actually want to go on a roller coaster. Like they've had enough.
- So I think the issue is one of cause and effect... So part of it is saying, yes, I agree with you that people are going to seek out that experience of trustingness. And then part of it is to say, as we are in the situation of being here, whether it's through law or technology design or policymaking, if you recognize that having situations in which trust is an important component and developing trust is an important component of it, you may choose to support different institutions. You might choose to stand behind different rulings around what platforms should be allowed to do and what mistakes should people be allowed to make on their own. And I think we will seek out those trust affirming and creating situations, but we are living in a world where, to a large extent, they are less of an inherent part of everyday life. Whereas, you know, in a much more rural and non-network non-computational setting it might've been a big part of everyday life all the time.

Charles Nesson

- First, just an observation on hitchhiking. I found it very interesting to listen to you, but my own, uh, connection with that experience has come mostly in traveling in other countries, uh, with my spouse in rental cars and having the experience of picking up hitchhikers. In terms of similarity, I'm with Wessel, in that it's the more dissimilar that made the pickups more attractive, and that's just an observation going by.
- But my question actually is about affective trust, and the connection with emotion and with the brain and with individual humanity, and your reference to the fact that you're going to deal with institutional trust in the next chapter. So I found myself wondering all the way through whether corporate institutions are capable in terms of affective trust, and curious about your contrast between the kind of trust that we as individuals experience, which you very well described, and the kind of trust that the institution engages in when dealing with institutions or with individuals. That's my question.

Judith Donath

- We had a little bit of a discussion about it early on, where just to reiterate what we said earlier, our trust in institutions often comes because we personify them in some way. Like, I don't think we trust, in this effective sense, a sort of vague, giant corporate entity.

Though people may trust, um, smokey the bear or Ronald McDonald to the extent that we personify them. The next chapter doesn't deal with institutional trust. It deals with the creation of entities, such as the grandchild of the sort of artificial entities that people have today, as they become more and more sophisticated at being able to appear trustworthy, but are actually computational systems and what that is going to do to our experience of trust.

Charles Nesson

- So for example, Harvard, take Harvard. Do I trust Harvard because I personify it?

Judith Donath

- I think there's different levels of trust in an institution like that. So yeah, in the really metasense, I would say it's not quite trust, but where I think you do get that kind of trust is in this, um, in that kind of group membership way. So if you go to alumni reunions at a university where you go to football games where people are like, you know, cheering their team on. I mean, if you look at where money comes into universities, not necessarily Harvard, but for certainly the big football universe, and it is true I think at Harvard and Yale also, the biggest alumni contributions are all around the big teams. So they are all around that kind of group forming. We are in a group against this outgroup of this other team we are playing against. And you know, if you look at all the sort of Yale, Harvard rivalry, there's this sense that the Yalies are on one team and the Harvard people are on the other. Even though they are very similar institutions with very similar people. That's where I would say these mechanisms of affect and loyalty come in. I mean, that's not true of institutions in general, but in a place like a university, you certainly get that at that type of in-group out-group group-based trust.

Charles Nesson

- Well, that's from the point of view of the donor, but from the point of view of the institution, when we're talking about institutional trust in individuals, is that affective in your...

Judith Donath

- You mean like, I'm not, are you talking about an institution trusting individuals or individuals trusting an institution?

Charles Nesson

- I'm talking about institutions trusting individuals.

Judith Donath

- Uh, well, I guess that would depend on your belief about an institution having an internal subjective affective experience, which is not how I think of institutions.

Charles Nesson

- Nor I, that's why I was asking the question.

Judith Donath

- I would say no, I would say that's a different use of the word trust.

Primavera de Filippi

- If we think about the reverse, I'm just thinking, the extent to which I do trust the Berkman Klein center, which I do, and what kind of trust that is? Because I think there is a large chunk that is effective trust. And I'm wondering what it is that I actually trust? Because I don't really trust the logo of the berkman center. I think it's the same with companies like Google or I don't know if I would trust that one, but, um, you know, Google, it's not about like, there's no personification that is made of the company. I think it's mostly the direction. Do I trust Twitter? Well, do I actually trust Jack and things like that. Do I trust Facebook, it's mostly about Mark. It's effective. In the case of Berkman, there are not many institutions that you feel you do have effective trust. Even though there's a lot of institutions that you might trust.
- And I think it's usually the people that you perceive as being in charge of that institution. The people that are actually managing and operating the system. So I'm not sure if it's always the mascots. The Twitter bot doesn't have any impact on my affective trust towards Twitter. So I think there is a distinction to be made between when you talk about trust towards an institution, oftentimes you can actually expand the trust and find the various actors, which are actually individuals, that you can actually express affective trust towards.

Matt Prewitt

- My dad said to me growing up, 'never get a tattoo of the company's logo', which is an expression of this idea, because the company doesn't care about you. I think it's interesting because the world is full of people trusting institutions that don't care about them and are potentially incapable of caring about them. It's interesting to consider whether this is like some sort of bizarre necrophilia or whether it facilitates lateral trust in interesting ways.
- For example, if I'm a fan of a sports team, then it solidifies my relationship with the other fans. So here's an example of something that happened to me recently and it's interesting. So my wife and I were going to meet some people that we didn't know to see whether we could adopt their dog. And the dog had a collar with the logo of the Oakland Raiders football team. And my wife said, 'Oh, he's a Raiders fan'. And the guy said, 'Oh, that's

great. I'm glad that he could go to a good family'... I'm actually a 49ers fan. I didn't say anything. And I was thinking when he said, 'Oh wow, he could go to a good family', I felt a very positive, affective trust. I felt good that he felt that way.

- And then it made me think, 'Maybe I should be more of a Raiders fan. Why do I, what do I have against them?' And that kind of thing is interesting, how our loyalties can actually move around in response to these affective attractions and repulsions. I just thought that's an interesting example.
- And then my question, if there is one, to Judith. I mean, to me, the big thing moving in the background of this is like, I totally agree with your definition of everything. I think that effective trust is the whole game. It's really what we want. And my question is like, can we try to build systems or institutions that cultivate it? Or is it just this slippery thing that we can never quite get our fingers on? Can we build spaces in which people can sort of explore their relationships with each other to build more trust or, is that just like a dangerous game where we create new monsters? That's the big thing I'm trying to figure out. And I'm really curious what you think.

Judith Donath

- I think the answer to that is yes. And I think the value of articulating affective trust as a goal is because it does involve that kind of risk and vulnerability. So if you don't have a good reason to say, well, we should let people take risks and we should let some people end up getting hurt or having some problems because of this. If you're just saying it and you don't have a reason for it, it doesn't seem reasonable. It seems more reasonable to say, well, we have the ability to get rid of this, why should we? But I think articulating it in this sense gives us the ability to argue why you want to leave risk and vulnerability in particular circumstances and why those actually are really valuable. And to be able to look at it in a way that you could explain why that's a goal. So I think, yeah, we certainly can build things like this. And I think it does have, you know, even immediate applications in terms of how you choose to regulate things, et cetera. So it's, it's not just a hypothetical, I think it's a very real sort of practical thing. And as for your example, did you get the dog? Did you end up getting the dog?

Matt Prewitt

- No, I turned out to be allergic to him.

Judith Donath

- Oh no! I'm sorry. I think it's a really nice example because it also shows that this isn't just like the simple thing where you can say, well, you know, here's where your trust is and here is where it is not. It's sort of a constant experiential thing.
- I would guess also that part of what happened here was, this discussion about the Raiders and everything was in the context of this dog, which I guess at the point that the whole

Raiders thing was happening, you hadn't yet discovered you were allergic to, and the dog had some of your affection. And it started being the space where shifting your loyalty a little bit was going to be perhaps the price of having this dog. The dog is really cute. I think in a different context, the Raiders thing wouldn't have seemed so appealing, but it was the dog that primed that

Matt Prewitt

- I think it was more of the person than the dog. I was sort of touched by the extension of faith.

Judith Donath

- Okay. Yeah, a lot of it is just very situational.

Primavera de Filippi

- If I can also try to answer Matt's question, because this is one of my favorite insights from the aggregation of the previous session. Judith put it explicitly in those terms in your paper: you say trust is a means to achieve confidence. In terms of why we would want to give space for trust, you put it in a positive sense, that we like the affective sense of trust, and therefore we want to give space for it, and if we have too many constraints, then there is no possibility in building further or reinforcing trust.
- I think there is also potentially a negative answer to that, which is that if the way that you want to achieve confidence is to increase constraints to the maximum, you will by consequence significantly reduce the agency that people have within that system. And so in that sense, trust is not something that we cherish. It's not something that we want to promote in and of itself, but it's kind of like a necessary cost that we decide to accept and to foster within the system, in order to make sure that we can reach this level of confidence as you call it, usually we call it cooperation, so that we can reach cooperative dynamics without eliminating completely the agency of people. So without relying on these extremely predefined sanctions and constraints systems. And so trust is there and we like trust, but not only because of the positive affective effect that trust has, but because it's the necessary correspondence of actually leaving agency to people and therefore vulnerability and therefore risk and therefore we need to build trust.
- If the choice is between pure constraints and sanction and no agency, or risk and vulnerability and agency and therefore trust. We would probably prefer trust in order to achieve cooperation. So I think both of them could be combined in order to justify why trust is actually a positive thing that we want to promote as a trigger or driver for cooperation.

Judith Donath

- I agree with what you're saying. I think that the relationship between agency and trust and risk are very closely tied in together. I would have to think how I would put the notion of agency into this, but you know, on the surface, I agree with you. But then agency becomes something that you have to sort of pin down to, and then you get into complicated questions of where are things deterministic and where is agency, questions that I was hoping to leave for somebody else.

Eric Alston

- I think we've mentioned several interesting examples of affect that I don't think as I understand your definition, so I'm not putting words in your mouth, but I don't feel like it actually satisfies the definition of affective trust, which seems, this discussion has made me realize, seems to hinge integrally on reciprocity and inner personality. And so I think trust in an institution or organization fails on important dimensions of the interpersonal reciprocal nature of trust and in the chat I called it sort of emotionally tinged faith in inanimate things, because in addition to say, you know, the corporation creating a mascot to have a stylized representation of the person that would actually create an affective trust between the individual and the organization. For me, that's always an inferior substitute because the mascot can't truly reciprocate the deep inner personal bonds that you evoke when describing affective trust.
- But I also was thinking of the deep sentimental attachment I have to a 15 year old pair of jeans. They're Japanese. They fit the definition of wabi-sabi and the underlying philosophical ramifications of that perfectly. I have a deep emotional connection to those jeans. And I like it, I trust them in interesting ways. But how on earth can they reciprocate in any meaningful sense? They can't, they're an inanimate object. And so I think what must be particularly frustrating in even having these types of discussions, let alone writing a book about them is when you evoke trust, for many people, you're evoking all of these things alongside it, that they use the word trust in their heads to mean, that it's closely related and even has an affective component, but actually isn't what you're describing.
- And so the TLDR of my, I guess, comment is the reciprocity component of affective trust to me seems to be doing a lot of ontological lifting in terms of defining the boundaries of what you consider to be trust and other affective relationships, including ones we have with inanimate objects or impersonal organizations out there in society.

Judith Donath

- I think a lot of this, again, just gets into the notion that in the end, all of this has to do with what we experience in our mind. So I think reciprocity is important in building up trust in terms of practically how you do it. But I think definitionally, it's not integral to it. You could have trust without reciprocal relations, particularly if you believe that there is that sort of reciprocity.

- In terms of things like mascots and the corporations, one thing I don't think is clear from this is that I need to do a better job of articulating the relationship between the trust that comes from group membership. That you're like me. I just agree to trust you because you're kind of like me or you're a type I trust, and the type of trust that's built up between individuals. And I think they are very similar pieces and they're related. They're not identical.
- So when we talk about trusting Berkman or something, a lot of that is the trust of feeling that the people who are in this organization are people like you. It's a little bit of the people who sign up for couch surfing, there's something like them. The organization is the equivalent of an ethnicity. And so it's not that your loyalty is specific to the organization. It's that the organization defines the group.
- And then with our trust in inanimate things, part of it is also that there aren't really clear boundaries where these things, you know, start and end. Some other people might just see something as, yeah, it's a corporate entity, whereas for others, you know, there's a huge amount of affection and loyalty to Disney. And all kinds of Disney characters and things, and people relate to them, you know, there's kids who have this great relationship with like, hello kitty. So it's hard to draw that line because a lot of the sense of reciprocity or your belief that the other would be there for you, has to do with what you imagine of them, which may not have that much bearing on reality.

Eric Alston

- I think that's spot on. Very tiny followup question that you don't even need to answer necessarily, but in a society with on average greater levels of trust among members, would you predict a greater or lesser need for mascots or corporate personas? And I keep going back and forth. I don't have a clear answer. So I'm not trying to lay a trap or anything.

Judith Donath

- I don't know. I would have to think about that, and it probably depends on what you mean by need and what their role is. Is it that that's a group where, you know, that sort of thing is looked at a lot versus you, you know, I can also make, uh, an argument. My partner at one point accused me of having what he called male answer syndrome, which was the ability to answer any question, no matter how little I knew about it, but so just on that one, say, I'm not sure.

Primavera de Filippi

- I actually have a hard time seeing mascots or whatever personification of institutions or corporations as something that will trigger trust. I can see it as some kind of maybe affection, just like your jeans, but I don't think you trust your genes because your genes have literally no agency. And in that understanding of trust, you can have no trust with

no agency. So I don't think you trust your jeans, you might, you might be confident in the fact that they will not break and whatever, but I don't think it's a matter of trust. Unless you're a kid, maybe, and you think that the mascot is actually alive. But if you just have a logo and you know that this logo has literally no agency, to me, I cannot conceive of trusting something which doesn't have agency.

- That's why, when I'm thinking about an organization, and when I think, is there affective trust towards this organization? I'm thinking, where is the agency? What is fueling and creating this operation? And that's what I might trust or distrust. But whatever cute little animal you might put on the face of the organization, it might create an emotional thing, just like I might like a little Teddy bear. Uh, but I don't think I would consider this to be anything close to trust at any level.

Judith Donath

- So another thing now I realize I need to be a little clearer about is it's not so much that this has to be reciprocal, but that we've been conflating... And this is my fault. Because both of these are effective elements. But part of the trust element is your being trustworthy. And that a lot of it is that I think the affective component of acting in a trustworthy way, the problem with the con man isn't that he doesn't trust you, it's that he doesn't feel a need to be trustworthy. Or the character in the brother's karamazov. Because of that lack of affect in that relationship, they did not act in a trustworthy way.
- So in terms of the jeans, it is one sided, but there is a trust level, in that Eric is going to behave in a trustworthy way towards those jeans. He's not gonna throw them out. He would actually be sad to see them sitting on a garbage can. He is going to be trustworthy and take care of them. And so that's a big piece of it. It's not where you get the important sort of kickstarter in the behavior, it is really in what makes people act trustworthy, and that affective component that makes people act in a trustworthy way is also really important, not just that you trust the other.
- Getting back to the mascots, I think some of that just varies by person. Like I am not interested in celebrity culture. I don't follow teams. I don't find mascots all that enticing, but I have to recognize that other people do. And those mascots do exist. People do show up wearing them. People collect plates from corporate entities that they display with pride. And again, I think Disney is a good example. A lot of things like that. You do see that type of loyalty. Both feel that they have to be loyal, that they shouldn't betray this company that they like by going to a different one. And that they do expect that kind of loyalty from them. It may be misguided, but I think that affectionate level can exist. It's just, not everyone feels that.

Primavera de Filippi

- I want to make a distinction. What you're saying here is actually, the extent to which people act trustworthily towards someone, so in that sense, you're not the trustee, you're

the trustee. So I can see clearly how Eric will always act trustworthily towards his jeans, but that's different from trusting them.

Judith Donath

- Exactly. It is very different, but it is part of the trust relationship. We can't only look at the side that's trusting.

Primavera de Filippi

- So if there is a trustor, then the fact that I act in a trustworthy manner might encourage the trustor to trust me, not necessarily, but it's probably a driver towards that. But I think that the reciprocity is often between the trustor and the truste. It is like, do I trust you to be trustworthy, but it's not. Do I trust you to trust me? And I feel there are lots of situations in life in which you might trust someone with your life, and that person will never trust you about anything, but I trust that person to be trustworthy. Even though I know this person doesn't trust me at all. So reciprocity is not the same thing.

Judith Donath

- Yeah. Or like just, I mean, both from a neurochemical standpoint and, you know, some of the strongest trust there is, is that children's trust in sort of, you know, unless they've been really betrayed by their parents, that their parents are literally the people who would die for them in many cases. So you can be a very untrustworthy child and not be particularly nice to your parents, but still have implicit trust that they will do everything for you. So you can definitely have very one-sided, extremely one-sided relationships that still involve a great deal of trust in, in their formation.

Primavera de Filippi

- Yeah. And, and I think there needs to be a type of reciprocity, but its trustworthiness towards trust.

Charles Nesson

- Yeah, I'm sorry that we lost Matt. I was completely taken by his observation that his interest is in trust building. How do we get a group to build trust? That I have to say is exactly my interest and particularly so it has been my interest in the context of teaching classes in the COVID environment. Going from a residential institution in which students are able to build trust relationships by all of the informal connections that they make in dining rooms and living situations and in between. And coming into a COVID classroom environment where that's really their only connection and where the focus is on doctrinal material.
- This issue also brings into play this question of agency, which is very pertinent to it. And my specific interest in the design of trust-building that can be built into combining a

face-to-face interactive environment amongst students with a pseudonymous, text-based environment. And it's precisely in the text-based pseudonymous environment that students are given complete agency, the agency to speak or not to speak, the agency to say what's on their mind without feeling like something's going to come back on them as having been inappropriate or stupid. And in an environment in which frankly, the classrooms have become dangerous places for students to speak out, especially on controversial issues, the opportunity to have a sandbox-like environment in which to discuss an issue pseudonymously before engaging and leading into the face-to-face environment is exactly in pursuit of a trust building objective. And one in which the example you were focusing on, Prima, with the increase of agency combined with the absence of sanction is exactly what it takes for students to take risks, more risks in what they say, and to learn about the sanctions that taking risks may entail when they actually engage in face-to-face within the full community environment. So I associate myself with Matt's observation that it's the trust building that counts.

Judith Donath

- I think this is all part of trust, but I think what you're doing with the sandbox is removing risk, because you're basically saying like the risk of doing this, whether it was online or offline, I think here, this is an intellectual risk. So I think the risk would be just as high as we've seen in our email. Our lamented email list. Um, maybe pretty high just in text, but so what you do is by separating people's identity, the huge risk of feeling humiliated, or, you know, just that kind of personal attack is removed. So the student can drop an idea into this place and see, you know, is this greeted with warmth? Is it torn apart and left to be eaten by sharks? So they can sort of try it on there and then make their choice whether they want to take that idea and have it associated with them in the high risk experience of an identified discussion.
- So in some ways what you're doing there is a little different because I think what's happening is that your pseudonymous sandbox is like a marketplace of cultural acceptability, where you get to see what the acceptability level of different concepts are, who accepts it and why? So when you have doubts about what the reactions to an idea would be, you can drop that idea in and it will get measured and reacted to, and then you can decide what you want to do with it.

Charles Nesson

- I agree completely, but I'd also say that it's an environment which functions with everyone knowing that their participation in it is required to be civil, if they're going to be able to maintain the existence of that space. And so as discussion goes forward, they learn both to value the community of open discourse and to assure themselves that they are indeed capable of talking about difficult subjects without causing damage to

themselves or others. And so in that sense, it's very much a community trust building experience, right?

Judith Donath

- But I think a lot of the community is because you have this parallel pseudonymous space that is parallel to a non pseudonymous space. That is a small group of people who know each other and are developing certain bonds among themselves. And so there is that knowledge when they're in that other space, that behavior in that space, while it wouldn't necessarily be reflected on them personally, that community is a commons that's of value to that group as a whole. But I think you also recognize in some ways it's a fairly delicate situation and involves a lot of boundary keeping around the classroom and building up that sense of community and the knowledge that you could have one adamantly disruptive student who could actually wreck it pretty thoroughly if they wanted to. So part of it is trust in the institution of the class as having created a boundary against that level of trust ending disruptor.

Primavera de Filippi

- And I think it goes back to the discussion we had with couchsurfing, which is by creating this sandbox, you're creating more possibility of agency. And therefore vulnerability. So you're, you're, you're facilitating the creation of this trust. Which is not necessarily interpersonal trust because I don't know who is talking, but it's trust within that tribe, within that specific community. And once the trust within that community has been established for the sandbox, then it might become easier to develop interpersonal relationships with the individuals themselves once you know their identity, because it's preconstructed. Just like it would be hard for me to just make a friend with anyone in the street. But if it's someone that comes from, you know, um, the Berkman center, then immediately I'm more keen to create a trust relationship because of this, uh, collective trust that I have in the community that it is associated with.

Judith Donath

- Yeah. I'm gonna try to phrase it a little differently. This will make you really happy Charlie. I think your pseudonymous space, um, while it's not risky in the same sense, what it does is for your classroom as a whole, because everyone knows that anyone could have anonymously acted disruptively in that space, but didn't. The classroom as a whole knows that everyone there is trustworthy because they've all been in a situation where somebody could have, without any repercussions to themselves, been very disruptive ,and they know that nobody did that.

Primavera de Filippi

- And I think that's a way of exploring indirectly what are the general and shared values. And so at some point you enter into a class, you don't know what the shared values are. And over time you start understanding what people think and therefore you understand, Okay I do have also encapsulated interests, and we have aligned interests, and therefore it becomes easier. So it's this complement to constructing trust at the collective level in order to make it easier to bootstrap trust at the individual level at the same time.

Eric Alston

- Yes. I have an alternative gloss to what's happening, which is rather than trust being built, a set of a set of risks or costs that are typically present in the personal classroom environment have been removed. And so I'm at a minimum closer to my law school days than professor Nesson, um, and many people were very hesitant to speak their minds in class because of the multiplicity of games that we were engaged in simultaneously. And I mean, games in the social sense. So some individuals were looking for their future partner. Some individuals were signaling ideology, some individuals were signaling intelligence, and it was all tied in an integral way to them personally.
- All of the risks and rewards of those games go out the window when you have a pseudonymous sandbox. And so suddenly this person thinking I'm an idiot, this person thinking I'm an evil conservative, this person thinking that none of those are present in the single game you've constructed. And so for me, it might still be a very effective way of building trust, especially in the ultimate classroom environment that people step into with a better idea of shared values as Primavera just noted. But I do think an important component of what's going on is also that a lot of the stakes that are present, probably even more so at a school like Harvard, a lot of the stakes that are present have been removed from the table in the narrow confines of the sandbox that you've constructed.

Primavera de Filippi

- I want to say this because I think we discussed it a lot ,and I'm just curious whether your take is the same or not. You actually use the notion of confidence in a very different way from the way in which we've talked about it until now in this reading group. You actually put confidence as a calculation of probabilities. Whereas up until now, we've been relying more on Luhmann's approach to confidence, being when you actually don't even take into account a probability calculation, because you're actually just confident that things act in a particular manner. And it's nice because actually most of the discussions we had until now was to which extent confidence can lead to trust.
- And you're actually talking about the opposite, to which extent trust can actually lead to increased confidence. So I'm just curious to see, like given all the discussions you actually attended, to which extent do you see this interplay? Because it depends again on the definition. One of the main guiding questions of this working group is actually how do we increase trust? And we've been looking a lot at confidence and its implication and

whether by building more assurance through more constraints and sanctions, uh, can we actually increase the degree of trust that we achieve in a particular system. And what you're saying here, by increasing trust, I actually increase the confidence in the system as well.

- But I feel like we are not necessarily talking about the same concept of confidence. And so what I'm trying to ask is like, given your understanding of what we define confidence to be, what are you actually referring to when you say constraints, sanctions, and trust are all ways of achieving confidence? What would be another way to say this sentence without using the word confidence?

Judith Donath

- That's a really good question. And I think I am struggling a lot with vocabulary through this whole very late book. Um, so confidence may indeed be the wrong word. Um, I mean the book itself is really about communication, and at what point you believe something enough to change your mind. And it could be at a really, really tiny level. It doesn't have to be a big, you know, change in what you think is true. Like just what lets you accept some claim or some claim that somebody makes about themselves. You know, if there's another student in your class, there's an implicit claim when they say something that they're claiming I'm smart enough to make the statement, do I accept that claim that they're that smart or not? You know, so it can be a little impression making things.
- Confidence for me is really that sense of, do I believe X? Do I believe that this claim is true? Do I believe that this person is going to act as they say they will do in the future? And it can be sort of amorphous. 'Belief' I didn't use because confidence seemed to be something more that you could have varying degrees in. And belief just seemed more like the result, that that was on and off. Um, whereas with confidence, the reason I liked that word was, you know, and I'm still, it's still a piece I'm struggling with, was because of that relationship to risk. That at some level to accept something, you have to have some amount of belief that it's true, but how much that belief has to be… And when I say it's a probability, it's not like a conscious calculation of a probability, but just that if the risk and consequences are fairly low, I may accept things pretty easily, even if my confidence isn't that high. If I pushed at it, I might take something at face value just because the risk of being wrong is low. And so that's what I was trying to get at with confidence. I don't know if that answers your question and it is, I know what I'm talking about. I don't know if I have the right word.

Charles Nesson

- We've had an experience at the Berkman center with our email list, in which it effectively blew up in a discourse that was sufficiently hostile so that people felt like they were being injured by it. And it's been shut down at this point. Uh, and so when you're speaking about confidence, the confidence that, uh, I'm feeling as an immediate need is

the confidence that other people won't speak in damaging ways, and confidence that when I speak other people won't feel damaged by it. Again it's the classroom situation. It's not a classroom, but it's effectively the same, where you're not talking about a situation that's open to the net and where trolls are a problem. You're talking about a problem within a group and about a group and how the group establishes norms of discourse that make it work.

- It comes into particular focus with the contrast between confidence that's established by having a system of rule and sanction, so that the stronger the sanction, the stronger the confidence that such a thing won't happen. As opposed to an environment which is more or less free of sanction and where the strength of the community adheres in the accepted norms, subscribed by each individual, within the group. To me, it's, it's a key to be working towards that second as the objective, rather than the first. I'm not in favor of a classroom that's surrounded by the strictest of sanctions when anyone misspeaks. I'd much rather have a classroom in which people within the environment are more or less acquainted with the calculus of damage that can be incurred by different ways of speaking, and choose to withhold or choose to be sensitive to the problems that come at the margins. Does that make sense?

Judith Donath

- Yeah, I mean, I think we're very much in agreement that the experience of confidence through trust is a much more pleasant experience than confidence through sanctions. Even though the behaviors may look the same.
- But I think Primavera is pushing to have a better definition of what I'm meaning by confidence in order to understand how those things relate. I would say that confidence is something you need. There's something that is asking you to in some way, change your thinking, even at a very tiny level, you know, your belief about another, there's something that's happening, what level of what I'm calling confidence is that you need to make that change. And do you have a better word, Primavera? I understand that it's not what you would like to see used for confidence. And I think that makes sense, but I don't have that better word.

Primavera de Filippi

- Yeah. I just borrowed the definition of Luhmann and now I adopted it too much, but for me, like constraints and sanctions leads to more confidence. Whereas trust is something different from confidence, which Charlie was saying. You can reach that thing, which you call confidence. And I would like to find a different word. I think maybe the closest thing would be like cooperative dynamics. Like you can reach a particular spot in which you're willing to cooperate with someone, but it's not just cooperating, it's actually unidirectional. It's kind of like relying on, I'm willing to rely on something either because I'm confident, meaning that I know that there are constraints or sanctions or just

the law of nature. So I'm confident that it's going to happen exactly as I expect. And therefore I rely on the fact that, you know, the sun goes up every day and that's fine. Um, I don't trust that the sun goes up everyday, but trust for me is a different category. For me these are qualitatively different. There are all the things increasing confidence, and all the things increasing trust. And the question is, are they complementary, are they supplementary, are they in opposition to each other? We don't know yet, but trust and confidence both can contribute to achieving this reliance on something. Maybe reliance is a good one.

Judith Donath

- Yeah, I don't think reliance... Maybe belief?
- If two people ask me to borrow my car for a week. Um, and then you say, Oh, well, bring it back. If someone I've known for a long time, I can just say, here's the keys. I trust you. I am confident that they're going to do that because I feel like they don't want to let me down, so they will do what they have to do to get it back to me. If it's a stranger, I am going to constrain, or, you know, in various… I want to get into definitions between what the differences between constraints and sanctions are. I might say fine. I need some kind of deposit from you. Or, you know, um, you have to leave your GPS on your phone all the time, and I want to be able to track where you are. And there's all kinds of ways I can constrain that because I don't have the trust, but in both cases, I have achieved sufficient confidence that they're going to do what they said and bring it back to me in a week that I have acted and lent it to them. But I've achieved that confidence through different means. And I don't know what other word to use instead of confidence there. But I'm open to suggestions.

Primavera de Filippi

- Yeah. It's a super interesting question because I think confidence and trust have so many definitions and, uh, they're being used very sloppily. As long as you define it in a particular way, it's fine. Like people will understand what you mean.
- It's just that by using confidence, you lose the opportunity of using confidence in the other sense of the term, which is when you have this certainty because of the lack of agency, um, which I think is just as interesting, and to be able to contrast them. But because you're defining trust as something that contributes to confidence, then it's difficult to understand, to which extent confidence is contributing to your definition of confidence.

Judith Donath

- But if I wanted to talk about a lack of agency, I would say that's constrained. Someone is constrained to act in a particular way. So a constraint makes it impossible for you to do the wrong thing or the unacceptable, the thing that's not allowed.

Primavera de Filippi

- Yeah. And sanction is kind of something in between, right? So with sanctions on the one hand, you have more assurance that someone will follow up, but you don't have full assurance. You need to trust something else, you need to trust the sanctioning authority that will actually act upon it. And so sanction in between. It doesn't destroy the agency completely. And I feel sanction actually is an interesting thing that requires or entails a degree of trust in the larger system.

Judith Donath

- Yeah, in the book there's a whole chapter on sanctioning that comes before the chapter on trust and talks a lot about religion and the surveillance of sanctioning belief in gods, you know, so it can be a real thing. It can be an imaginary thing, you know, and it's there. It's important too, because the technologies of surveillance are changing so rapidly. And so the cost of sanctioning is going down.
- But you're correct in that if you have to line these things up, both conceptually, but also historically. That line that you made that said constraints are here, sanctions are here, and trust is here, I think you can make developmentally. Because the book was about signaling and, you know, starting from, uh, animal behavior. And you can have signals without any kind of society, and there's constraints and costly signals, you know, in ants and insects, viruses, et cetera. So that's a very simple version and you start building up sanctioning in much lower animals and trust comes at a much higher level. You actually have to have real social institutions to have trust. So while they operate on very different levels, they certainly emerge in that order too.

Primavera de Filippi

- And I feel like the contextual questions are super interesting. With Balazs we're trying to write an article about trust in context. She's actually looking at those concentric and eccentric relationships in which constraints, sanctions, and confidence generate trust, and how trust in the higher system generates consequences in the lower system. And I think you, you cannot really talk about just confidence and trust without understanding what is the outer system that generates this trust, or that generates confidence and then the interplay within both.

Judith Donath

- Thank you so much for this opportunity, Primavera, and also for the initial discussions that got me working on this whole piece to begin with. It was all because of you. So I am tremendously in your debt. Thank you so much. I can't thank you enough.

May 6 — Coeckelbergh (2012): Can We Trust Robots?

Attendants: Juan Ortiz, Mark Coeckelbergh, Divya Siddarth, Brett Frischmann, Matt Prewitt, Charles Nesson, Liav Orgad, Ori Freiman, Victoria Lemieux, Primavera De Filippi, Judtih Donath, Eric Alston, Morshed Mannan, Wessel Reijers, Jack Henderson

Key concepts:

- This paper asks if the question of trust is applicable to robots.
- The phenomenological-social approach to trust puts less emphasis on individual choice and control than the contractarian-individualist approach. It defines trust not as something that needs to be 'produced' but that is already there, in the social.
- In so far as robots are already part of the society and part of us, we trust them as we are already related to them. While robots are neither human nor mere tools, we have sufficient functional, agency-based, appearance-based, social-relational, and existential criteria left to talk about, and evaluate, trust in robots.
- Cultural differences impact how we think of trust in robots: all criteria depend to some degree on the culture in which one lives and that therefore to evaluate trust in robots we have to attend to, and understand, differences in cultural attitude towards technology (and robots in particular) and, more generally, cultural differences in ways of seeing and ways of doing.
- When it comes to shaping conditions under which humans can trust robots, fine-tuning human expectations and robotic appearances is advisable.

Transcript:

Juan Ortiz

- I think the paper starts with this breakdown of what do we mean when we say we trust a person? And so it kind of tries to define this by opposition. So it creates these two groups, the contractarian-individualist and the phenomenological-social. And so we could say that the contractarian individualist starts from the idea that there are individuals who are in this kind of vacant space. And so they eventually see each other, engage with each other and establish a relationship. And through that relation, they build trust. And so the phenomenological social is a critique to this and starts by saying, well, actually, trust is the social bond and the social bond is trust. And the social defines the relation and the social also defines the individual. So we can think of it as going from the bottom up to some extent. And so, you know, when Mark was asking, who are you reading? And I guess what was interesting about this paper for all of us is that we have been reading the people who are to some extent criticizing or building with. And so in the paper, the

contractarians are discussed in terms of Luhmann, who says there is a coordination problem. It's about complexity and uncertainty. Gambetta who talks about trust as a calculation of probability with which an agent has to assess that another agent will perform a particular action and focus on Trust. And so we'll go into this in one of the next slides. But basically, it's not exactly as the previous two, but it's still based on the idea of this contract or an individual's idea and the rationality components that underpin it.

- So what is this phenomenological perspective? It kind of stems from a critique of the contract area and says, you know, what about the way in which children trust friends and parents? There seems to be something missing. An effective dimension of the contract sharing perspective is not capable of identifying. So trust is not something that needs to be produced, but something that is already there so that the paper and the rationalization might happen. But it's after the fact, right? That's something that we humans like to do to explain what we have done. But it's not necessarily the way we approach trust itself. And social responsibility is not the flipside of trust. It's actually built into the social and communitarian relation which crucially has nonverbal and implicit aspects of these nonverbal and implicit aspects of something that I think is it would be interesting to discuss as well as preconditions of trust. The paper presents to us as well, especially for the contract area. And there is this ability to use language, right? This is the key building block for the contractor and constructivists. Individuals rely on language to build trust to get to know each other. Then there is freedom and uncertainty. And so as part of the definition of trust, we need to have that the receiver can misuse or diverge from the trust that has been given to him or her, but also the social relations. And this is key for the phenomenological social point of view where trust, talk and talk about individual freedom presupposes these social relations. So what does this mean in a translation into the realm of robots? Right. And the paper is clear that when it talks about robots, it's something more than mere tools. Right. It's not something that merely executes but has something that is different there. And so from the contract Korean perspective, and this is where Therriault comes in with iTrust conceptualizes humans and robots as agents, so it kind of and synthesizes the idea of an agent so that it can encompass both, even if the shared social and moral norms are not present. And from the phenomenological perspective, the focus is not on what they are, what they appear to be. Right. And so you can acknowledge that robots don't have language in the way that humans can interact with language or freedom, and yet we often treat them as if they did. So perhaps this gives space for us to talk about something like Quassey Trust and how we would assess trust in robots. So perhaps from the contractor perspective, the easiest way would be can the robot do what it is supposed to or expected to do? And from the phenomenologists perspective, we might say, well, did that grow into a social relation? Is it felt as another robot perceived or interacted with as if it were another? What I thought was most interesting was that it also causes us to take a broader outlook of our relation with robots, as that is, do they help us to understand ourselves and how we shape ourselves? And the

last section is focused on the idea of culture and how culture is kind of trans versus all of these questions because of the phenomenological perspective. And so the key role that it has to play. And so in this sense, it's a much richer approach than the idea of these synthesized agents. And so in terms of discussion, one of the things that I was most interested in thinking about was I like this idea of trust being kind of embedded within the community and perhaps inseparable from that. But perhaps when we think about communities of people as being bound by mutual need, and I think that that's why humans kind of our social and effective communication often takes place through expressive emotions like a baby that cries, not knowing what crying means, but through evolutionary processes that cry kind of triggers things in the parent that understands that something needs to be done and it's urgent. And so we develop a culture of trust and duty. And so I wonder if that is something that robots can engage with when you think of robots, they're designed to be kind of self-sufficient. It's not that the robot is hungry and therefore expresses a need or is sad and so it can engage in that kind of set of relationships. That is, of course, bidirectional. And so I wonder how that might evolve. And with that, I was curious to know your thoughts on some of these topics.

Mark Coeckelbergh

- Just a quick comment. I like to end the mentioning of emotions that need something more effective and things related with the body and so on. So that's, I think, also quite an interesting area to talk about trust in relation to that, in relation to needs, in relation to emotions.

Divya Siddarth

- As we think about the concept of trustlessness, which is something that we started with, it's interesting how the phenomenological social account kind of precludes trustlessness even being a possibility, because it's situated us all in these kinds of preexisting trust relationships and looks at those as social. And so, you know, the individual contractualist approach may allow for a trustless interaction, I think that's something we've been going back and forth on. But it seems that this social approach doesn't, and that to me feels intuitively like a better encapsulation of what trust looks like in online and offline systems, although I'm not sure exactly how to articulate it beyond precluding the need for trustlessness.

Primavera De Filippi

- So what you're saying is that according to this specific definition, a trustless system is basically starting from the premise that, because you don't have any social relationship, then there can be no trust that will ever emerge from that. So just just to just to inform Mark about the long discussions that we have on this topic, one of the underlying

question of inquiry that we are trying to have is if in a particular system, mostly technological but not only, we managed to increase confidence, so providing guarantees that things will happen in a particular manner and so forth, is the provision of those guarantees and thus increased confidence in the system capable of, I mean, what was the impact on trust? And one possible way of thinking about that is that, by increasing confidence and by increasing technological guarantees, we are removing the opportunities for trust to emerge because we are reducing the amount of freedom, we are reducing the agency. And therefore those relationships of trust cannot be built. On the other hand, the other provision is perhaps sometimes trust would be established, but it is not established because there is too much uncertainty, and so increasing the confidence in the system might enable the creation and the establishment of trust relationships that wouldn't happen otherwise. And we're going a lot back and forth and trying to explore those different definitions.

Brett Frischmann

- I got stuck on the first premise, which is that robots are necessarily more than just tools. And so the reason I'm asking it this way is just getting stuck on why robots and why are they more than just tools? I think you can apply all of your analysis even if they are just tools. Right. So taking the epistemological approach to thinking about robots, if they're just tools to focus on the makers and the builders and the constructors and the users and deployers of robots as tools. So I think much of what you say about robots would still hold even if we didn't say that robots are more than just tools, but it might focus on different, slightly different things. So I kept thinking why robots are different from bridges or other sorts of infrastructure? Because the functional difference is they have a degree of autonomy to act self sufficiently, or is it about how we perceive robots to be different and so we treat them as differently. And that leads to a series of observations about trust in robots that is different from trust in the power grid or trust in the traffic management system on the roads or trust in the legal system or whatever. Like we can go through a whole bunch of other aspects of our built environment that we regularly trust, which I think is different than analyzing trust as a component or characteristic of a social relationship that we have that makes sense.

Mark Coeckelbergh

- I think it's partly true that some of the things I say can be said without making a claim. But I think there is something special about privacy here, because to the extent that they are perceived to be more than tools to be like a child or an animal, for example, they then tap into the way we are more socially wired, the way we relate to others. And in that sense, they become part of a sort of, you know, social network as it is perceived and experienced by us. And that creates then this kind of, you know, experience of the robot as the social actor that can be trusted, cannot be trusted. So, yeah, I think in that sense, it's

more than a tool. It's not just about the function of the robot, but it is the way it is; it relies on this kind of social web that's already there as experienced by the people.

Brett Frischmann

- Is that not an aspect of the design of robots themselves? So we could build other things that would also leverage our wiring to develop social relationships and things that we relate to in certain kinds of ways. So you could think of a stuffed animal. So you said start with children and how they feel about their parents. We could also think about stuffed animals, actually figurines and other things that people relate to a little bit differently. Not that those things are any different than just a tool or a device for play, or a thing. But they're designed in part to trigger reactions and relationships and things like that. And I'm not saying that those reaction relationships aren't real. They're real and significant and they may lead us to develop something that matters. And then we want to talk about why it matters. But I don't know if it renders the toy or the stuffed animal or the bridge or whatever, anything other than a tool. Anyway, I did get stuck on that a lot of times. But, you know, I thought that's super interesting.

Mark Coeckelbergh

- Yes. For me, it's gradual. So the puppets already, you know, evoke this whole social world much more and tap into this kind of social and emotional stuff. But, um, I think the robot is scary because it has some wisdom here that we're talking about the robot who has a higher degree of autonomy and interactivity especially than the puppet. And I think that makes this kind of mechanism stronger so that this whole idea is about trust and so on gets really going. Whereas, you know, with the puppet, it's more a projection that we do on puppets versus the robots speak back or do some things. So I think that's phenomenologically a different kind of thing. And therefore, yeah, we are talking about this kind of trust question in the first place, I think whereas it doesn't come up with a hammer that it's like more trust in the sense of reliability, you know, is it reliable if the head of the hammer, if it comes off, comes off, then then, you know, it's not reliable. But we would have used the word trust in that sense. Whereas if the robot listens to my conversation and then goes to tell my brother or sister what I said, you know, some secrets are what I'm imagining. This children's situation, that's that's a problem. And it is because we see the robot as part of us, part of that social world. And that's, of course, often intended by the designer and of the robot. And we also get all these problems then.

Primavera De Filippi

- This is sort of like a discussion we had at the end of the last session, which is to me, as you say Mark, there needs to be some kind of agency, there needs to be some kind of freedom. And therefore there needs to be some kind of risk and decision to put yourself

in this vulnerable position in order to trust as opposed to just being confident or less confident about the tools. But to me, the question then will be like, when we're talking about trusting robots, I was talking. So the question is, are we always talking about confidence or trust? And you claim this is trust. But then the follow up question will be, are we talking about trust in the robot? Or are we actually talking about trust in the operator, the design of the developer? The human or the moral entity that controls that particular robot, so forth and so on? Yes, I want to make sure that that is not going to spy on me and and tell my secrets to my neighbors. But is it really trusting Alexa, or am I actually trusting Amazon.

Mark Coeckelbergh

- If it's put in these terms, it's really human for me. So we trust in humans, but because the robot gives this impression, we start, you know, saying the word trust in connection with the robot. But by doing this, we sometimes forget that well, first of all, you know, it is the robot. It's not a human. And second, that there is behind the robot, there are human designers and the human, the humans who make profits from, you know, selling this robot from getting the data about us, from the robot. So I think it needs to be seen in this whole kind of ecosystem where the robot is part of what generates trust. But if you ask me, like, you know, should we trust robots? Well, no, because trust is about humans. But I try to understand that people perceive the robots as, you know, somebody they can trust or not.

Liav Orgad

- I love the paper and I kind of disagree about these things. So I think in a way we can think of trust in the Aristotelian way of the way he thought of virtue. So basically you are virtuous if you fulfill your function. SO if trust is about fulfilling the function, then these functions can be naturally made and like humans of certain functions and emotional functions. But also it can be artificial. It can be amended. Right. So in a way these robots would be. But I can also think of trust in this regard as kind of artificial, manmade, whether we trust democratic institutions to fulfill their functions that we provide them, whether we trust God, if we created gods to fulfill the functions that God is supposed to create. And in this regard, I don't think that trust is about humans, trustees, about the dialog of of humans with themselves, not necessarily with another entity. To answer the question whether these things are natural made or or artificial, then fulfill the functions that we hope that each would fulfill. And in this regard, that's what I was taught when I was hoping, like in the paper. I think it's great because you give some pre-conditions. But then ideally we would start with the question, what are the conditions that provide the kind of the requisite precondition for trust? And then and this is very much related to the question whether it's to then ask the question, OK, what's different? Do we have a different case with regards to robots? And you kind of presume that robots are a different

case. And I am very much with Granth in this regard, that it's almost if you kind of take it to be self-evident that they are not mere tools. But that's not so clear. Why not? So, so, so in a way, whether I was finishing is just you could you could build like a two stages trust the regulatory mechanism in which one we talk about the prerequisites for trust in things as a dialog among ourselves, kind of fulfilling the function, whether it's a thing that you automate men and then ask the question of where the robots present a different case.

Mark Coeckelbergh

- I think it's a good exercise. You're proposing here to talk to tease out the differences. But yeah, I do make the assumption that the robot is a smart tool, because if it was not like that, I think it doesn't make sense at all to talk about trust. A robot is not a tool. I mean, not in the sense that I know for sure that the robot falls into a certain ontological category, but more we experience this kind of situation. The robot is more than a tool. And I think that's the basis –

Liav Orgad

- WHat's the difference between robots and God in this regard? When asked, can we trust God or can we trust democratic institutions? Then you would also classify them as mere tools? And if so, based on what distinction?

Mark Coeckelbergh

- These are huge questions. I think in the case of God, we probably in Christianity see God as a person. So there because we experienced as a person, we get the same kind of thing, like with the little robots that we talk about trust. And so regardless of the question, what God really is. Yeah, and the same with the robot. We don't really know what it is with democracy. It's more complicated than democracy is an institution. I would also call it the kind of institutional tool or something, a social tool that we have. So it is a kind of technology, but it's, of course, also a technology that's very much connected with human beings. And it's not one thing that pretends to be a human or something. It's an institution. It has to do with sweets, all kinds of norms and rituals. And it's interwoven with human beings. So it's quite, I mean, a complex thing to think about. What is it to trust in democracy? So I'm not sure I can answer that, but. But yeah. It's a different thing than I am then having this one other, whether it's an artificial order or a divine order or something, because they're there, we just you know, we easily incorporate within our, you know, our kind of social ontology in a sense of. Yes, this kind of structural thing that's already there. And we try to apply the same kind of rules and we see them in the same kind of way because we're used to that among humans.

Wessel Reijers

- I just had two small questions. One of them is about whether one needs to thematize temporality, in order to get an understanding of trust. Because here again it is like the main distinction is a temporal distinction. It's about like the already and then not yet, between the two different approaches. So I was wondering, like, to what extent we actually need to delve more into the question of temporality to understand this distinction better?
- And the follow up question is about the characterization of the phenomenological social. Which in a way, it's kind of a contradiction in terms because like at least in the classical phenomenological positions like Rousseau and Heidegger, they actually also have a very individualistic approach. And of course, they talk about the life world and worldliness. But for example, in Heidegger, eventually the distinction between authenticity and inauthenticity leads to authentic experience in the individual. Or in a kind of solitary experience of being towards death. So that's social. So it's like I was questioning some of what there are like approaches that do try to bridge this gap between this very individualistic phenomenological approach and social reality. But the gap is not self-evidently bridged sso to say. I was wondering how, you know, how we actually get from this, you know, from the phenomenological approach, which is actually very individualistic to this and the social approach which in a way is not self-evident.

Mark Coeckelbergh

- Yeah, great question. So to start with the second one, I think we definitely need a social phenomenology and that is kind of what I'm suggesting in this kind of paper. So that's why I'm not really a big fan of just the smart individualist focus on the you know, that the kind of ego or something like that was it was early on, the kind of phenomenology or early was so and so. I think phenomenology can help here to get us focused on the experience of the human and, you know, the relation to the object. But I think we need to go towards a more social approach because otherwise we're doing a kind of mirror thing to what's this age and kind of view. That's right. Starting from the individual and then trying to build up the social. And so what I'm saying in a paper to society is already there. That's why I like it so much with the two arrows. So we need this kind of account of the social that's already there and from which in a way, the individuals are made to emerge rather than the other way around. And whether phenomenology can pull this off or not, I think we definitely could use more different approaches there. And some maybe in hermeneutics, for example, there could be other approaches for the, um, the temporality. Yeah, I agree with you. This is something that needs to be developed. Absolutely. I think we in general, in thinking about technology, don't pay much attention to that. And, you know, apart from some people that I do get from people who are going to do this, I think there 's not much work on this. I think there's a lot of work needed on this particular topic. And yeah, also in relation to trust.

Eric Alston

- I actually kind of began reading the piece in the camp that's already been well articulated surrounding, is this just a tool? What's going on here? Because I'm an institutionalist. I teach at University of Colorado, Boulder. And so my toolkit for understanding the world is one where the three overarching ontological categories are people, natural resources and rules. But I really enjoy this piece because it made me realize that the distinction between robots and sort of what we call personal interactions or people interactions, for lack of a better word, is blurrier than we might give it credit to be if we say this is a tool and this is a person. But first, I have a very short clarifying question. Is a robot for you necessarily something existing in physical space that approaches human attributes? In the sense that or would it still qualify for your definition if I put on a VR headset and interact with a very compelling human avatar, is that no longer is that sufficiently outside that category you're describing as a robot? Because I kind of see as an integral component to anything being a robot as opposed to a statue as being the set of directional processes that that thing has, whether you want to call them algorithms, whether you want to call it A.I., but a key component for a robot to generate anything approaching an emotional response like trust is it needs that set of processes. I think it's a little silly to talk about trusting an inert statue that just has a lot of very compelling human features. So would your definition of robot encompass, say, the virtual avatar I'm describing if I put on a VR headset?

Mark Coeckelbergh

- I would not say it's a robot if it's virtual, just to stay within the conventional definition of robots, most of us would not say that's a robot. But I agree with you there. There's a lot of similarity between the two. So here I think your question would be needed to tease out the differences between being situated in the physical space and in this virtual world. What I think is pretty strong in physical space is that it comes quite close to the way we deal with other people, which are also situated in physical space. So, again, I think it's a gradual thing that the avatar does almost everything that the robot does, except it doesn't, you know, especially does things that're even the same. But the feature of being situated there in the physical space is not there. So that creates possibly a different experience, a stronger experience, not that different. It's a difference in degree, but it's it's a different experience. But I don't I don't know the exact answer to that. So I think it's a question that is not completely solved yet. What is this difference? What does it mean really to have this physical thing there as opposed to not having the physical thing there? So this is just me, you know, continuing on the line that I've been doing. But I'm not I'm not sure what the differences are.

Eric Alston

- I make that example to say I think these things are separable. And indeed, the definition of a robot might well involve a physical component. But my argument is, it needs some type of compelling algorithmic component to also surpass the threshold of just being a statue. And so with respect to the vertical of the algorithmic component, developing a set of processes that seem complex enough, I think the boundary between human interaction and that type of interaction is actually blurring in really profound ways in that I'm seeing increasingly the semi automation of interactional processes.
- Four examples that kind of to me emphasize this continuum. If you think about people who deliberately have a persona on specific platforms, video streamers, I'm friends with professional DJs and they have a very specific persona that they enter into when projecting on a specific platform. Some of that's live streaming, which is much more of an interpersonal connection. But increasingly, if it's on Instagram, the format is constrained, it's predefined, it's prepackaged, and everyone has the same interactional format. And so I'm seeing the semi automation of interactional processes. The second example is a frustrating interaction I had with somebody trying to verify that my best R.A. worked for me and this was somebody in a different country reading off a list of questions that were dichotomous. And the question that really got me was, are they eligible for rehire, which usually indicates, you know, a very bad thing. In my reality, they were ineligible for rehire because they graduated and were no longer students. But if they became a student, I would hire them back in a heartbeat. But that person could not answer that question. They were almost like a robot in terms of how rigidly defined the interactional process was in that type of process is increasingly being done, at least in first instance, by a chat. And so I contacted my bank in a chat bot answer. I contacted FedEx and a chatbot answered. Eventually I can trigger the right set of responses to connect to a human. But that is not the only chat bot I've encountered. I've also been messaged on Instagram by Sexy Juleanna three seven four saying, Hey, please come to my private website. I found you of all people and I want to talk to you boring academic man. And I am beautiful. And I'm like, I, I think you're a chatbot, but one with very different motives linked back in an interesting way to a set of humans. But in every instance I was thinking of what's interesting. The link between the chappe is that it's a link back to a set of people and their incentives, it is like a social context. I see it as an organization with a sexy chat bot, probably a criminal organization that at some point wants my credit card information or bank details. The bank that is the bank organization that it's linked to and potentially linked to a human. But to me, the set is like we're automating our interactional processes and compressing the scope in which they occur and all the triggers that they can then create. Sometimes they eventually lead to a real person, but they always link back to this social context, this organization whose incentives I have to think about in order to understand whether I truly trust the chat.

Mark Coeckelbergh

- Yeah, yeah, absolutely. And I think the system also discourages us from doing that. So it's kind of an anchor that you're just interacting with Chabert and that you don't ask these questions, that you don't ask the question, who is behind it? What do they want from you? Why did they do that? But one of your examples struck me that the people themselves start to become more robotic and so elected that just say like it's not just the problem, it's not just a step up, but it's the whole interaction that gets an ultimatum. And that's something I'm worried about, too. And it shows that the robot in that sense is, you know, reshaping social interaction between people and in general, you know, between anything or anyone and making it more one dimensional, to use a term from our Coosa. So that, I think, is important. Also, I would say that robots are not humans, but they are human in the sense that they're humans behind the development of robots. And so for a critical perspective on robotics, I think we absolutely need to reveal that human world behind that and before it and after.

Eric Alston

- And just to put a final point, linking it back to the true definition of robots is I think we are possibly semi automating the physical interactional components as well. I think the immediate value to society is less clear than automating these interactional processes because most of the innovation seems to be derivative from the sex doll industry. But nonetheless, there is deep interest in automating or developing very real life, like intimate interactions by Ikara.

Mark Coeckelbergh

- If we have more automation technology, I'm worried that we will start speaking the way that the machines understand the stuff we start moving, the way the machines understand us. Imagine a security situation and a machine by image recognition codes calculates how much probability there is that you're that that you're problematic for security, you know, and people will adapt and move like robots through the process in order to be seen as normal. So there's a kind of normalization and as you would say, a semi automation there of the interaction and of the behaviors. And that's really worrying. So I'm also like you. Are you worried about what happens to the humans rather than what happens to the robots? And so my way of thinking about the moral status of robots, for example, has also always been about, you know, being worried about human ethics really.

Eric Alston

- The clear example that springs to mind of us moving in order to please the robots. I think of the position I've assumed one hundred times for the scanner at the airport.

Primavera De Filippi

- There's this beautiful irony where we used to ask robots to try and prove that they're actually human or that they're not a robot. And now it's like us humans constantly with CAPTCHAs trying to prove to the robots that we're actually human.

Victoria Lemieux

- What I would really want to drill down on is a question that Wessell raised, which was the relationship between the individual perception and the social world. And I guess it kind of relates to this concept of our perceptions and perceived social reality versus if you accept this premise, an objective social reality that exists. So my question is a bit of a thought experiment: if a robot is designed to keep secrets, and that is one of its features of trustworthiness or the features in which we can place trust. And then if the robot is used by a bad actor to conduct criminal acts, and the robot behaves in a trustworthy fashion, the bad actor can trust that robot to keep the secrets. So we can say, you know, that there's this trusting relationship between the bad actor and the robot and the robot is trustworthy. But then the robots being used in a way that from a social perspective, is very untrustworthy. And so from that social perspective, we would say we don't trust that robot. That robot is, you know, basically behaving in a way that is engaging in this bad, you know, the bad practice. So I guess what I'm asking is for your thoughts, Mark, on how do we resolve that tension that might exist between, you know, this individual's trust in the robot and. Pretty for the robot to be hit, you know, behave in a trustworthy fashion and yet a kind of a broader social perspective from the robot, which might resolve to. The robot is very untrustworthy. Yeah, I'm just really interested in your thoughts on that. And I have in the back of my mind, you know, a situation that pertains to what the group is looking at with blockchain technology. But I don't want to muddy the waters with what I'm thinking. I just want your reaction to that thought experiment.

Mark Coeckelbergh

- Interesting question. Yeah. I mean, one could say that there are different levels of sociality. There's an interaction itself and the relationship of just one person to another person. But then there are sort of larger social holes. And I think that's also in sociology, you know, you would distinguish between different levels and different kinds of social realities. So I don't think that's so much a problem. But you're right that there can be tension. I think that's very well observed. There we really need to ask a question like trustworthy for whom? Because it could be that also different groups have an advantage from a particular use, whereas other groups miss out. So I think there's a kind of larger level indeed and a level where we can ask political questions about trust. But I don't have a theory worked out to link the two or to spell out all the different levels here. But this is definitely true. Yeah. Something that needs to be looked at at different levels, different social entities.

Victoria Lemieux

- I see that tension, too. And the reason I asked and motivated to ask is I struggle with it in thinking about block chain technology in the sense that we have a technology that's designed to have properties of trustworthiness, and yet it is used by hackers like thinking of it being perceived as being very untrustworthy or associated with untrustworthy. And so that tension and I haven't quite figured out how to resolve it either. So I was hoping –

Mark Coeckelbergh

- I'm afraid to resolve it. I mean, I would say the transaction there is trustworthy. You know, I can trust the transaction because it's, you know, with this crypto stuff, it's really trustworthy to me as an individual. But then from a societal point of view, you know, what have I done with the transaction? I think to distinguish at least between these two different things helps us to see more clearly in this discussion about blockchain, because I think there's a lot of confusion about that.

Judith Donath

- I think perhaps a useful point is that I enjoy the paper and I think a useful way of thinking about what distinguishes robots, particularly in terms of trust, has to do less with their actual capabilities or construction and a lot to do with how we perceive them. Because people can think of a car that they've driven a lot, especially before the era of modern, highly electric cars, as something trustworthy and perceive it to have its own agency. So I think the issue of trust can often be seen as a subjective thing that we apply to things that we perceive as having agency. And that's where you can have two machines that do the same task. You know, whether it be, you know, to unload your dishwasher or take care of your aging parent, one may be designed to look exactly like a machine and. It's just sort of, you know, moving something down the hallway for you and you might not think of trust with it, or to the extent you do, you're aware of it as a machine. You think of the fact that it was built by engineers and, you know, are they competent? What they do, what often can transform something and why we call it a robot is it presents itself as some kind of being with the implied agency, which then one keeps us from thinking as much about the engineers because we think about that agency as being internal to it. And I think, I mean, this now is bringing my own concerns. But I think a lot of what then becomes particularly problematic is that a lot of the signals that we use to make sense of the trustworthiness of other people in particular, but it could be animals, et cetera, has to do with things that are in some ways inherently tied to how actually trustworthy they are, whereas it's much easier to manufacture a robot that creates that impression of trustworthiness. But it's completely separate, unrelated to whether it is actually trustworthy. So its mission may be to, like, secretly gather up all the information I can about you and sell it on the black market. But it will be designed with big eyes and

a sweet voice. And it can do all kinds of things that we as humans, most of us would be pretty unable to do because it's hard to be that deceptive, whereas the design of the robot can be incredibly disarming. And so that's where a lot of the problems in trustworthiness versus the appearance of trustworthiness in the machine come in.

Mark Coeckelbergh

- That's right. Yes. The designers and the people who employ the robot may, you know, abuse our, you know, our abilities to trust them the way we trust them for a purpose. So that is true. And I think the example of the cars was also good to see that it's also to do with subjective stuff about work and also culture. I'm very sure that some people would say that they would really trust a purely mechanical and electric car more than the electronic kind of computers on wheels that are now making, you know. So I think that may not have anything to do with the actual performance.

Judith Donath

- There's an interesting article about our beliefs in the autonomy of things, and that something that acts in unexpected ways can often make it seem more autonomous. So a car that likes every time you start, it just starts off and it drives and it makes the same sound. It does everything totally reliably, it will seem far less autonomous than one where you're like, oh, you know, I have to jiggle it in this way to get it to work. So, I mean, I think that's a lot of why mechanical, particular mechanical cars may seem more in that state of trust because we perceive them as being more agent-like. If you look at the strategies that people have used when they are designing robots for the Loebner Prize competition, the one that's meant to, can you pass the Turing test? A lot of it is about putting in, you know, deliberately designing mistakes and things like that into their dialog that it's clearly robotic when it has no mistakes. But you pretend that it had typing errors or clauses in it, et cetera, to make it seem something more lifelike. And if you look at the design of most robotic pieces, they do have a lot of features that are just designed to be trustworthy, even by people who aren't trying to do anything nefarious, just like people will trust it more.

Primavera De Filippi

- Is the fact that we can trust robots because there's this kind of social relationship that needs to be established? And is it the case that we can only trust a robot if we can feel that this creature has the ability to trust us in return? As opposed to, can the robot trust us, and is that a factor that will increase the trust that we could give to that robot? Instead of starting from the social relationship first. So if I could create a robot that was giving me the feeling that it trusts me, would that create a particular social relationship with that robot that will increase my chances of trusting the robot back?

Mark Coeckelbergh

- Yes, I think that's right. So I will trust the robot more if I have the impression that the robot also trusts me and has expectations about me and so on. So the more I see the robot as a social actor itself, the more I will also trust or mistrust them if something goes wrong. So I think we have in our social relations this kind of assumption that there's a reciprocity that I trust the other and the other trusts me. So that definitely I think can increase the trust if we get that impression.
- If we actually go from the social relationship perspective, whenever you see people that put themselves into a situation of vulnerability, and it relates to what Pettit is also saying in terms of actually giving trust to someone is promoting the fact that this person will act trustworthily. But also if I see kids or whatever that puts themselves in a position of vulnerability towards me, I'm most likely going to also assume I can trust them, not that I can rely on them, but I can trust them as individuals.

Ori Freiman

- This distinction between the social phenomenological approach to trust and the individual contractarian is such a powerful distinction that I wonder what happened to it?
- In the field of philosophy of technology, the continental approach is quite a bit removed from the individualized-contractarian approach. The major approaches, for example, are phenomenology or actor network theory or critical theory of technology. They rely much more on Marxism and post-phenomenology. Than analytic philosophy of technology, who uses trust in the sense of the contractarian discussions. And I think that that is maybe a bit of a feature of the field as such. There might be interesting crossovers in the field of speech act theory. That's where often the two approaches meet, because I know that hermeneutics has some interest in speech act theory, which is also often invoked on the contractarian side. So I think that that's maybe a potential candidate for overlap.

Primavera De Filippi

- I'm still a little bit confused about this question of: why are we talking about trusting robots? As opposed to trusting whoever is the individual or collective entity that is responsible for the actions of that robot? There is the physicality and the AI-ness of the robot. And I guess there is a point in which the AI-ness is strong enough that you can no longer associate the behavior of the robot with the institution. And then there is perhaps like a cat. And then maybe at this point it makes sense to talk about the robot as trusting in the AI, but as long as it's just like a physical body that is mechanical and is still operated and controlled by a particula operator or entity, I find it very difficult to talk about trusting robots.

Judith Donath

- I don't think you have to even invoke any kind of AI to get into the subjectivity of people's relationships with objects that appear to be some kind of life-like entity, if you like. One of the examples I've used is Tamagotchi, where, you know, these are these little pocket games where you have this thing that was sort of like a pocket chicken. And so you're supposed to keep it alive. And people went to enormous effort to keep these things alive. Like at some level they knew that it's not actually dying. But you had that responsibility as your pet and you took it on. So you had that kind of emotional response to it. And so I think what you're saying makes sense from a very, very rational point of view, that as a completely rational actor, of course, it really likes where the issue of trust really lies is with the manufacturers and the people who design it and cause it to act in particular ways. It's true, but in terms of what people's experience of trust has to do with the experience of being with it. And so I think they're both really interesting questions, but I think that's the reason why we talk a lot about these issues of trusting the thing, because that's the interface where the trust actually happens or not. I think Eric's example is interesting, even if you don't care about robots at all, it helps you have that complicated line that goes from. AI robots to really simple ones. And then you get to people like doing telephone scripts and where you have people at various levels enacting scripts. And you can go to the Goffman level that says, well, that's our whole daily life is enacting some kind of script. So there's this really interesting continuum that you get in terms of where do you get trust, where do you get autonomy? What's the relationship of any of these things to the script or the society or the ties behind it?

Primavera De Filippi

- I have a hard time considering that I'm trusting a Tamagotchi as opposed to I might care for a Tamagotchi, but I can’t put myself in any vulnerable position toward it.

Judith Donath

- I wasn't saying you necessarily trust it. I was using it as an example of how simple an object starts to elicit that type of emotional response. I mean, there's very little in the design of the Tamagotchi that asks you to trust it. But I think you can extrapolate that there are pretty simple mechanisms you could design that start to elicit a sense of trustingness.

Juan Ortiz

- I think that's part of the challenge of the phenomenological kind of approach in that it's quite useful in providing a descriptive account of things. And sometimes it, in my opinion, feels a bit like a cop out because it says, you know, you're rationalizing after the

fact when you engaged in that relationship, you weren't rationalizing it, you were just living through it, and now you're trying to theorize about it after the fact. And I think the challenge comes when you try to be a bit more normative, right, with some of the questions that you're posing Prima. And I think that many of us struggle with, like if this is not only descriptive but normative and trust is relevant, once you invest a lot of resources in making something trustworthy or seem trustworthy or gain your affection like the Tamagotchi, even if you know, for all societal purposes, it's not really worthwhile for you to spend perhaps hours on a Tamagotchi or trusting some machine that actually has no agency or or shouldn't be trusted as an autonomous thing that is worth your trust. But I think that's where the phenomenological kind of approach runs a bit short or gets tricky when you try to move from the descriptive to the normative of how things should be. I think for the purposes of the paper, he's being very clear that it's descriptive and he's working on robots and perhaps, you know, the expectations a bit like he was wondering. I think the expectations maybe in 2012 were different than the ones we have now. And I think now the focus is A.I. isn't as smart as we expected it to be. And there's humans behind it and they're biased and so on.

Eric Alston

- I think there's something interesting in the Tamagotchi example, and I took this point to be an emotional connection or a bond is born whether or not you would call it trust. But the implication that the thing relies on you for sustenance and survival and is designed on some margins feels like a life thing. But what's missing is risk, at least risk for you. And so I think risk and uncertainty are, at least I see as potentially an integral component of trust, but not sufficient, maybe a necessary component. What do I mean by that? I think the other thing as being, it's human made, and to evoke not likely the recent tragedy in Mexico of the Metro Bridge collapse that killed last, I saw over 20 people. I would say in an interesting sense they were trusting the infrastructure, which was the fruit of human construction. This afternoon, I'm going to go into the mountains and scramble sort of free climb on boulders for a good portion of the afternoon, which could well kill me. I'm not trusting the mountain. I'm trusting the bridges. I will cross as I get up there to not collapse because they are the fruit of human construction, preventing me from potentially life threatening risk. But it's very weird for me to say I trust this thing. That's it. That could well kill me. The mountain. I don't trust the mountain. I may trust my own decisions to not go.

Primavera De Filippi

- I would not say you're really trusting the bridge. I would say that you are confident in the fact that the bridge will hold because you trust the engineers and you trust the supervising authorities and the municipality to have done its job in constructing the bridge. But there is no actual relationship of trust that goes into the bridge. You’re not

even questioning if you trust the bridge. If there is any question to be asked, it's do I trust that the people involved in the construction of that bridge did a good job? And I think that's where the distinction between confidence and trust comes about. And I think if we take the bridge and we transpose it to the question of robots, as long as I can see a direct connection between the developer, manufacturer, operator of the robot and the actual operation of the robot, then I trust the operator. And therefore I'm confident in the fact that the robot will act as planned. But I'm curious to explore: I think there is a point in which, even if I do trust that the operator and the developer and so forth did a good job, I also know that there is a lot of contingencies and uncertainty and like I'm in my self driving cars and I never know what might happen and how it might make a decision. And even if I do trust those actors because of the increased amount of agency that has been given. And of course, we are not talking about General AI, but as we get closer to that, then there is a point in which I might find perhaps myself at least talking about trust to a robot, and I'm wondering like, what is this threshold? When is it that something without the preconditions for my perception of the trust relationship to be cut, from being only based on the operator and just being confident in the product? And actually because of agency, because of risk. Because of autonomy, with regard to the operator, all of a sudden I need to also, in addition to the operator, I also need to trust the system itself.

Judith Donath

- About the Tamagotchi again, when we talk about the trust relationship, I think we tend to forget, probably because it's also less relevant in terms of Bitcoin, that there's also the question about when are we trustworthy? And that's where the Tamagotchi comes in. If I start to make a cup of coffee and then I'm in a hurry and I leave and I shut the stove off and the coffee never boiled, I don't feel guilty that I didn't make that cup of coffee, that I've somehow let my coffee pot down. Yeah, I just didn't make the coffee. But the experience with the Tamagotchi is premised on the idea that it can trust you, that you believe that you are worthy of its trust. And what made it so compelling to people is it's not just they didn't want it to die. I think they felt really guilty. They didn't want to let it down. And so when we talk about these relationships with robots, part of what's also important is to recognize the extent to which they elicit our desire to behave in what we perceive to be trustworthy ways to them, which can also be quite manipulative.

Morshed Mannan

- I've been thinking about the DAO Model Law project in the sense of, how can we use this insight in terms of designing liability rules? Just like how Primavera was talking a little while ago about trying to find this distinction between, when we are trusting the system where we have, let's say, the person who designed an AI or a robot and we are trying to see whether we trust them and then whether that allows us to be confident versus, you know, whether this distinction is not clear anymore. I'm curious as to, you

know, sort of like extreme or edge cases with this, where let's say you have an AI or a robot as a director of a company. And obviously this is an example that I have written about already. And I'm curious as to how, you know, this perception of trust in this robot who has this particular role, whether the fact that they have a different position than, let's say, a Tamagotchi or even a healthcare robot, has implications for the liability they have. And, you know, in the article, he talks a little bit about this when he talks about responsibility, and the question of whether, along with being trusted, whether they can hold the responsibility. And conversely, if they are also working with other directors, in this example, there's also a question of whether liability rules can be designed in respect to the others that they're working with. And it's also referred to in the article about trust between agents as well as trust between humans and agents. So it's also the question about whether the robot is also reposing trust in others. And I think this becomes really important in terms of thinking that whether when we say we are reposing trust in someone, what are the material consequences of this responsibility as well? Does this mean that the robot should be able to have, for instance, a certain amount of wealth to be trusted or an insurance or some other type of way that a liability rule would be effective upon it? So that is one of the things that sort of struck me and made me think about how this would be different than, let's say, a bridge or another example this sort of liability would attach to someone else, not on the bridge itself.

Primavera De Filippi

- The question of insurance is a very good point. I probably would be more likely to interact with a self driving car that is insured against whatever damages. But to me, that doesn't mean I'm going to trust this car. That is actually a very rational probability calculation of to which extent am I willing to take the risk and what is the proportion of this risk and what is the amount of insurance and and only if that qualifies within my risk range, I'm going to trust this car. I'm not sure if I would call it trust, because, again, I think it's a pure probability calculation because I don't feel I'm putting myself at the mercy of the car. It's more a matter of how confident I am that there will be damage, what's the probability, and given that assessment, I will decide. There is a risk but it does not depend on the trustworthiness of the actor or agent I want to engage with. So while I see and it's oftentimes suggested as an interesting solution to promote interactions with robots that might not be trusted, that is a solution to facilitate interaction, but I'm not sure that the solution is triggering a relationship of trust.

Morshed Mannan

- This discussion about things like insurance or other types of remedies, I think it also, in a way, relates to this discussion that he has in the paper between the sort of phenomenological approach and the more contractarian approach, I guess, in the sense that I feel a lot of these discussions about, you know, this sort of liability where AI is used

by a corporation, does take this contractarian approach. And I was trying to think of how this more social phenomenological approach would be applicable or useful in this context. And maybe, as you say, like whether it incentivizes action or not. Yes, I agree with you. It does incentivize action. I wonder maybe, you know, this predisposition to trust which he talks about in his article, whether the existence of insurance helps, and because you know, that if you have reliable insurance system, et cetera, like there are certain preconditions for this, you know that you would be compensated if there is harm caused by this, in the same way that, you know, the existence of compulsory motor vehicle insurance can help enhance confidence and driving on a highway. Whether this would also not only encourage interactions, but also in a way create the conditions that, you know, Mark talks about in his article that make us gravitate towards trusting because of the fact that it floats in the system as something that's always there as opposed to other types of liability related sanctions like damages or something else that you would have to go for separately as a way of dealing with this. So that was how I would see it being related to the predisposition towards trusting.

Primavera De Filippi

- I think what Kate is talking about is similar to what Judith was talking about, which is sometimes it's irrational and you know it's irrational and yet you still develop a social bond and sometimes it's irrational, and yet you still develop a trust relationship. And I guess if we want to go into the normative, I guess it's actually a very interesting manipulative tool in which if I'm not trustworthy as an operator and if people were to assess my degree of trustworthiness. They might not want to interact with my devices, but if I create little devices which are cute and vulnerable, maybe they disconnect the trust link and they start trusting the device independent of the fact that they don't trust me. And then they are more likely to interact with something that they wouldn't if they couldn't trust the device itself.

Eric Alston

- Under your definition, is it possible to trust an object?

Primavera De Filippi

- I'm trying to figure it out, my obvious answer would be no. I think you can only have confidence in an object because of the lack of agency. One of the preconditions is the trustee needs to have a realm of freedom to betray your trust. And if an object doesn't have that, it's not betraying your trust. If the object is enslaved by an operator and it's doing exactly what the operator's doing. I'm actually vulnerable, and there is agency. But the agency is not in the device itself. It's in the operator that is controlling it. And yet, I think I don't know yet, but I think that there might be a point in which if the device

acquires a sufficient degree of agency, which is autonomous from whatever operator is controlling it, then maybe that's a possibility. Then maybe I might need to trust that thing.

Eric Alston

- I see one vector as being certainly the extent of agency that that thing can have, even if it's somewhat constrained by artifice in the design of the original operators. But I also think attenuation from the set of people, if that set of operators becomes sufficiently nebulous. What if they're long dead? Is it exclusively confidence that you have? I think you might say yes.

Primavera De Filippi

- I can trust someone that is dead. I can trust the intention of someone that was alive and now is dead. If a device was created by someone that I trusted and that person dies, I still trust the fact that it was developed with good intentions and that there is no attempt at betraying my trust.

Eric Alston

- and so the specific example I have in mind for this is I got engaged to my wife on the wall of an abandoned cathedral in the jungle, in the hoca in Mexico. And so we had to walk 40 feet along a crumbling brick wall to the end for sunset over this canyon. But what was certain is that the people who constructed the cathedral were long dead. I definitely didn't even not even didn't meet them. It's like I don't even have an understanding of, like building codes in 19th century Mexico in a remote area. And so one possibility is that it's just confidence. You had confidence that the wall was well enough constructed that you wouldn't that you wouldn't fall to your death. And it's so fine. But I have a second example, which is I think I trust my forerunner, my Toyota forerunner. And I have confidence in it, but I have an effective bond with my forerunner. I have almost died in it several times. It has, I take it, off-roading in extremely dangerous terrain and it is always over-performed. And I maintained it incredibly well and I maintain a highly personal relationship with my mechanic. And so there's a set of things that are mechanical with it that I know intimately as a result and part of my trust in my mechanic. But I also have a bond with this machine that it will prevail in highly risk related scenarios, that it will pull through like it has in the past. And indeed, I hope it will for the next three hundred thousand miles that I put on. And so I guess I see it possible for you to vest trust in an object, perhaps irrationally, very irrationally because of the critique you're making. But I think people can and do trust in objects that separate from this confidence that's linked to your trust in either a person or an organization.

Primavera De Filippi

- I want to make a distinction between the two examples. I think the first example does not convince me, because the first example, I think you do trust someone. And even if it's not the people that you trust, whoever is responsible for the tourists actually accessing the wall. You didn't trust the wall or have a social bond, you had confidence because it's been standing all this time.
- The second one is very interesting because the second one sounds like they've definitely a special bond with the car. And that's great. I think lots of people have special bonds with their cars or other objects. It's just a matter of what you call it. But if I was your psychologist and you asked me, I trust my car with my life, I would ask you what was the brand and do you trust the brand? But what is interesting, and what Judith was thinking about, is that when you do have a special bond, then it becomes an easy shortcut to refer to trusting the object when you're in fact trusting the manufacturer of the car.

Eric Alston

- But I trust it as a result of our shared experiences.

Primavera De Filippi

- But that's confidence. Confidence is about when you don't even ask yourself questions any more because of your previous interaction with a thing. I have a ton of confidence in the sun waking up every day on the east side of the world. I don't trust it, I just know it happens because I see it every day. So if I keep walking on a bridge and it never falls down, I might build confidence in the fact that it's not going to fall down in the next hundred years. But I don't trust it because there is no capacity, because I don't perceive the capacity. And again, I could indeed trust an object if I perceive the capacity of that object to betray me. And if the betrayal comes from the object as opposed to from the operator that is holding the object, then I might find myself in a situation in which I'm like, yeah, I do actually trust or don't trust that particular device.

Eric Alston

- So would it be fair to say that you see many people confusing a blend of affect and confidence for trust?

Primavera De Filippi

- Yeah, we often talk about trust in a very loose way to say a lot of things. And that's why we're very keen to make this distinction between confidence and trust, because it's actually a useful distinction to understand what's the mental process that goes behind it.

Eric Alston

- And I would agree with you that many people ascribe the label of trust to what is actually a blend of affect, some type of affective relationship and confidence. Now, notwithstanding that, I'd like to plant a humble flag in the possibility for trust in an object, perhaps irrationally so. I do think that there's a lot of confusion out there, especially surrounding the definition of this word and people misascribing it to something that it isn't, which is often confidence plus an emotion. But I don't know if I'm 100 percent convinced that it's impossible to trust an object

Primavera De Filippi

- I think lots of people might actually do it. But my hypothesis would be that if they do really trust an object, it means that they have so much anthropomorphized and personalized that object that they do think that this object could betray them and that the object constantly decides not to betray them. Your car is really nice. And even though your car could explode at any moment. It decides, no, I'm going to keep Eric alive. And if you do perceive that, even if irrational, then yes, you might be trusting your car.

Eric Alston

- Well, survey captains and how they feel about their ships. The feeling that I get when I make it off of a back country road. Having thought this car is not going to survive this, I think it is sometimes a similar feeling to how a captain feels following a massive storm or hurricane, which is, I think, a response that somewhat outstrips a calculus of whether or not the ship made it through. I mean, I've certainly seen a lot of narratives to that effect from captains, where it's like they have a deep affective relationship that they would call trust. I don't know if it's something we would call trust.

Morshed Mannan

- I'm curious what your thoughts are about, when they name let's say a ship as a female name and then, you know, seems to have a lot of both I think an effective component, as well as perhaps a trust component with the ship, whether this is necessarily anthropomorphizing it, or whether even the analogy that they have in mind is like with a horse or some other sort of animal that has a cognitive capacity of some sort, and that the analogy to the ship is being drawn to is like with a prized horse that the captain has in their sort of mental frame. Because I think it's a similar question to, you know, whether we can trust our dog, trust our cat, trust our horse. And then at what stage do we go from that to speaking about, you know, being able to trust a ship, for instance. Is it possible for the horse to betray us in that sense? Or do we give a different explanation about this? Because, for instance, with the Hardin terminology, maybe it's not possible

for a horse to encapsulate our interest or something like that. I'm curious as to whether it's necessarily in the ship example or even maybe the car example, whether the trust is using this mental model of a human rather than some other type of creature.

- And the other thing that I thought was really interesting with your car example is how it made me think about this example of this curious sort of creature of some common law jurisdictions of the purpose-trust and how this particular type of trust is unique in the sense that instead of having a beneficiary who is an identifiable person, they have a trust that is created for the benefit of like a cat, for the uptake of a cat, or that someone will die and that the trustees will take care of their cars when they passed away. And it made me think that I'm really curious about the history of this particular type of instrument. But it seems to me like a really interesting one in the context of this discussion, because it raises the question about whether this sort of trust construction could be used for the benefit of, let's say, maintaining a robot or your car, Eric, in the long run. And then taking it a step further, whether the trustee in such a trust can also be another artificial intelligence system, right now a trustee would have to be a human being or institution, but maybe that is also something that will come into being in the future. And thus, you know, they would be brought into a trust in a different way, even in the legal sense, rather than maybe in this sort of conceptual sense.

Eric Alston

- This is why I reference corporate personhood, Morshed, because part of the reason you have a trust with a purpose for the life of a cat is a cat doesn't have legal personhood that the property can attach to in most jurisdictions for their sustenance. And so a simple question is, is a river being granted legal personhood in New Zealand? For the record, I'm very sympathetic to at least if in my heart of hearts, if I'm being parsimonious about my definition of trust, it requires agency that can betray you. There are things we have a sufficiently effective relationship with that maybe I'm willing to call it trust. But forgetting that, forgetting Eric's forerunner that saved him on many occasions, I do think that there's a non-trivial ontological problem, which is I like the definition of confidence as in something that doesn't have the agency to decide whether or not to betray you, like I'm all on board with that. But the devil's advocate in me says, is there a problem when everyone out there is saying, I trust my ship, I trust my forerunner, I trust my cat. And we're like, no, no, no, no, no. That's not actually trust. Is there a problem there? Maybe there isn't. It could just be an ontological quibble, but at a minimum, a high proportion of people are using the term trust, and probably at least in some operational sense in their head also labeling it as trust. And so to me, it's interesting that it's like it's a pretty structural critique, or an analogous structural critique would be many people think they're in love, but they're not. That might be true. But that's big. So a lot of people are mistaken out there. They think they're in love. But actually, I'm calling it something else.

Ori Freiman

- It's different from the common sense language we all use when we go to the supermarket, when we speak with family, friends. But here it's, you know, a specialized discussion about trust that there's no choice but to take it to a level that does not exist in our usual language.

Primavera De Filippi

- Eric, I think your point is really good, assuming that we didn't have three months, if not more, of researching the various definitions of trust and trying to pick one that makes sense. And when I'm saying, you're not trusting it, you're confident, because we have pre-defined our own definition of trust and confidence. Now that we've got to focus on legitimacy, it's quite different. The notion of legitimacy is something that we can define at the macro level. And then it's inherently subjective. And at least, what we're trying to do with trust is to identify the useful definitions, and we found this distinction between trust and confidence very useful. And that's why when you say you trust the car, I'm not telling you that you don't trust the car. I'm telling you, according to our own definition of trust and confidence, what you're referring to is most likely falling within the category of confidence. The question is not whether you're trusting or not. The question is, given our definition of trust, do you trust your car? And I think that's interesting because if you do really trust your car in light of our definition, of course, then I'm very curious to understand what are the criteria and what are the conditions that are such that you do actually trust this car?

Eric Alston

- It's not the manufacturer. It's irrational. It's me ascribing that there is a set of random possibilities that could happen during any one of these risky trail runs in the vehicle. And so that set of random possibilities. It's almost like, I'm not anthropomorphizing, I'm lacking the right word, but it's like they become embedded into the vehicle that has the possibility of failing or not. And so the closest analogy I can come to is that of a captain and a ship, which is yes, it can be looked over in port. Yes, they can have trust in the thing that many people constructed or inspected. And they can have trust in the sailors who are on the ship. But there's a component of probabilistic instantiation of risk that almost becomes embedded in the thing itself. And effectively, it's like I'm hoping I'm trusting that this will make it up the mountain this time. And there's a deep affective component to that as well, in part because it's probably quite irrational, which is like vesting those risks in the material object that is most closely tied to the things you care about or the realization of those risks.
- Although it's possible that what I'm describing now is something akin to faith… What I like about confidence is it does tend to involve rational risk assessment. Will this hold,

given the things I don't want to happen. And past that maybe the right word for what I'm describing in my forerunners is deep faith. It's not trust, but it does have a deep affective component.

Morshed Mannan

- That reminds me of what Liav said about trusting God, maybe it's partly based on experience and partly irrational.

May 20 — Sumpf (2019): System Trust

Attendants: Ori Freiman, Eric Alston, Michael Heidt, Paula Berman, Primavera De Filippi, Wessel Reijers, Patrick Sumpf, Morshed Mannan, Balazs Bodo, Brett Frischmann, Matt Prewitt, Juan Ortiz, Georgy Ishmaev, Philemon Poux

Discussant: Ori Freiman (link to the presentation)

The journey begins with a vital distinction between the categories of 'System Trust', 'Confidence', and 'Familiarity' (in the introduction). That's part of the motivation that we will soon see.

2.1. The Dualism of System Trust
The section begins with presenting various debates about the concept of System Trust, with two examples for debates about how to correctly categorize the term 'system trust'. For example, is it more like **trust in collective entities** or more like **trust in persons**? Is 'system trust' based on **calculated decisions** or is it **compulsory trust**? (I read the term 'compulsory trust' as 'involuntary trust', but I am aware that it also involves a description of the system rather than the concept of trust).

Luhmann argues that in the term confidence unlike the term trust - there's no risk-taking or decision-making involved. The reason is some kind of an assumption that Luhmann has, which Sumpf rejects, that the influence of individuals on social systems is very low. So according to Luhmann's view there's no risk-taking or decision-making in systems because systems are not influenced by people. Therefore Luhmann replaced the term 'system trust' with 'confidence'.
Sumpf 's motivation is that he doesn't want to replace the term 'system trust' with 'confidence'. His whole point is that they are not equal – not the same. Sumpf breaks up with Luhmann also on the question of **genuine trust** - and siding with other scholars- arguing that it is possible to have genuine trust in systems. This idea heavily builds on actors participating in the system.

We are presented with the dualism of system trust as (my words):

> Type A: expectations that the system will work (Sumpf asks us to think about the function of abstract systems - like money or democratic elections)
> Type B: Where every trustor has different specific expectations (think about individual voters – with their individualistic decision-making, risk perception, and evaluation).

Despite some similarities, 'system Trust' does not equal 'confidence'.

2.2. Decision-Making and Compulsion.
Is 'system trust' a calculated decision (and a matter of choice) or is it compulsive trust? With the distinction that is made between decision-making and compulsion – the argument is that social systems are influenced by decision-making (and that's contrary to Luhmann's view from before). Therefore, the distinction that others made between trust as a choice and trust as compulsion fails to capture the essence of systems that have human influence in them.

2.3. Intersections in Trust Research
The aim is to challenge the assumption that general trust and trust in systems are that different. They have things in common:

1. both can entail a positive expectation for the future
2. both are directed somewhere: general trust is directed towards people or organizations, and the object of trust in systems is... systems.

For the rest of the chapter, Sumpf introduces what he calls "three essential components of trust":

1. Non-Knowledge: Trusting instead of knowing (key driver of trust)
2. Suspension: Suspending doubt and disbelief while reaching a state of trust / Transforms perceived risk into imagined certainty.
3. Expectations: Expected actions and intentions of trustees are reasons for trust / There are different levels that these expectations can be directed at.

In addition to the three essential components, here are three more:

4. Risk
5. Complexity
6. Control.

Together These six components are the center of attention of the book.

2.4. An Architecture of Trust: First Sketch
We are back to the topic of expectations. Following many trust scholars - trust can be seen as an expectation for something. Therefore the study of expectations is significant for the study of trust. Expectations can be directed at persons, roles, values, and programs (where programs is a term that represents a way to regulate the behavior of others… think about policy making for example). In this chapter there is a valuable table that shows that the architecture of trust is a result of having multiple objects of trust, with different expectations from them, about different functions.

2.5. Reflexivity, Construction, Attribution
Sumpf distinguishes between 'trust' and 'confidence' again. 'Trust' involves risk and vulnerability, especially when the familiarity with the object – a person is low. Trust leaves the 'future open'. 'Confidence', on the other hand, means that there's high familiarity with the object, which can be about anything, and since things work in a predictable way – he calls this a 'closed future' (maybe thinking of open and closed futures as expected and unexpected futures). 'Trust' and 'Confidence' both depend here on familiarity - But Sumpfs criticism is that nobody talks about what familiarity is!

Trustors who trust in systems change how they reflect on the system, and the trust relations change over time. This is what he calls attributions and expectations – that change. In his words they are fluid and reconstructive.

To sum up:

1. 'System trust' has a dualistic nature (general/specific functions). Can be action based on decision-making and be genuine.
2. A system is not merely a causal system that results with compulsory experience. It also entails a social component.
3. Recognizes six 'major components' of trust: non-knowledge, expectations, and suspension, risk, complexity, and control.
4. We need to study expectations. They range in their scale – from concrete to abstract, and in their objects (persons, values, roles, programs).
5. System trust relies on relations that change over time (Trustor's reflexivity, construction, attribution).

Conclusion: Sumpf wishes to advance the discussion on 'system trust' and have some empirical take on it. Therefore, he wishes to distinguish 'system trust' from other forms of trust such as (confidence, familiarity).

This is an inspiring text and I was left with a taste to read the rest of the book and with many questions of clarification. To end this presentation, I suggest two discussing points that we can begin with:

1. the view of "systems" as merely causal-mechanical, without social dimensions or any human influence (in the context of blockchains & smart contracts)
2. the architecture of trust - Components vs Architecture. What Sumpf referred to as "the six major components" – are characteristics of the concept of trust; maybe 'architecture' can refer to something more structural, where new phenomena of trust (that we did not have before) - result from the structure between entities and their relations.

Key concepts:

- System trust does not equal confidence as it can be an action based on trustor decision-making. In fact, trustors establish service expectations toward the general and specific functioning of a system, which are both phenomena of 'genuine' trust.

- The central condition for system trust to make sense is its connection to risk and decision-making: in cases where no choices are involved, there can be no trust. Experiencing the contingencies of social systems as something to be influenced by decision-making, on the other hand, leaves open the possibility for trusting or distrusting them.

- Current empirical developments challenge the assumption that system trust is merely a compulsory experience and underscore its possible decision-making component.

- Trust in systems generally can be treated like trust in persons: it can be regarded as a positive expectation for the future, but directed at a systemic object instead of a person or an organization.

- The basis of all trust research is the study of expectations. Based on established sociological concepts, service expectations toward both the general and specific functioning of a system can be ordered on a scale from concrete to abstract, such as persons, roles, programs, and values. The differentiation of these service expectations and their evaluation through assessment of system outputs by trustors form the basis of system trust and its research.

- System trust relies on attributions to different objects, self or other, for reassurance of trust and sanctions in case of disappointments. Attributions highlight the present-relatedness of trust because they can change over time.

Transcript

Primavera De Filippi

- I think we've been going back and forth about when something's trust or confidence. I enjoyed at the end of your chapter where you express the idea that this is a very exogenous way of actually defining trust and confidence, and to have something that we can observe externally as being confidence, or we can assume as an observer to the confidence actually is trust, or that something that we might actually perceive as trust might qualify as confidence, and that we cannot assess this without assessing what is the thinking and experience of the individual. That's something that we did not address yet within our thinking.
- In terms of how it applies, the exact same system might be used as a result of confidence for some people that actually understand the system, or it could just be trust by people who do not actually understand the system. It's the information that comes around it, as opposed to just, what is the governance structure that is designed within the system itself. So to me, that was a really interesting insight.

Ori Freiman

- Part One is the dualism of system trust. And there are debates about how to categorize the concept of system trust. So is it more like trust in collective entities or trust in persons? And is system trust based on the calculated decision, or is it something that you have no choice, meaning compulsory trust. Now as a side note, as I read the term compulsory trust, I wasn't familiar with it, and thought maybe it is involuntary trust.
- The motivation for distinguishing between system trust and confidence: Luhmann argues that in confidence, unlike trust, there is no risk-taking or decision-making involved. Now the reason is some kind of assumption that Luhmann has, which is that the influence of individuals on software systems is very low. So according to Luhmann's

view, there's no risk-taking because systems are not influenced by people, there's no decision making, so he replaces the term system trust with confidence. Now Patrick's motivation is that he doesn't want to replace the term, because they are not the same. And he thinks that there is trust in systems and therefore the system trust cannot be equal to confidence. So that's the motivation.

- Now Luhmann rejects the idea of genuine trust in systems. And here's another side note. In effect, Sumpf doesn't say explicitly what he means for genuine trust in systems. And I have reason to believe it might be different for different scholars. So that's the side note. Other scholars have no doubt that there can be genuine trust in systems, and siding with other scholars on this issue, Sumpf argues that it is possible to have genuine trust in systems. And the idea of genuine trust in systems heavily relies on actors participating in the system.
- So what is the dualism in system trust? System trust is not properly trusting that they work, but trusting because of how they work, firstly, because of the inner workings. So we are presented with two types of systems and these are my words.
- Type A: expectations that the system will work. Patrick asked us to think about functions of systems like money or democratic elections, and their main function. Each actor has different specific expectations, and think about those individual voters, for example, with individualistic decision making and perception of risk. And so system trust does not equal confidence. And that is the dualistic nature of trust in systems.
- Okay. So two forms of system trust. Is it a calculated decision? Meaning is trust a matter of choice or is it compulsive trust? And here is another side note: while reading this, I had the feeling that there is some kind of confusion in the level of analysis. Now, not sure if this is for my own limitations, but I thought that maybe the confusion is between the concept of trust and the processes of the system.
- With this distinction between decision-making and compulsion, social systems are influenced by decision making, and that is contrary to Luhmann. So the distinction made between trust as a choice and trust as compulsion fails to capture the essence of the system that has human influence in that. Okay. So having cleared a stage from traditional assumptions, between genuine trust and confidence in the first section, and compulsory and choice in the second section, we now move to intersections.
- Here, the aim is to challenge the assumptions about general trust and trust in systems. They do have things in common, for example, both entail positive expectations for the future. But the object of trust in general trust is a people or organization and the object of trust in systems is a kind of system.
- There are three essential components of trust. So the first category, he calls non-knowledge. Lack of knowledge is a driver in the formation of trust in that respect. The other central components are suspension, meaning suspending doubts, and expectation, which will be discussed in the next section. And in addition to these three essential components, introduces three more, in total six components, and these are the

components that are the center of attention of the book. So moving through the fourth section, so expectations from before. Following many trust scholars, trust can be seen as expectations for something. And there's a valuable table that presents this, and it shows trust as a result of having multiple objects of trust with different expectations from them, but different functions.

- And moving to the last section. Patrick distinguishes between trust and confidence. Trust involves some kind of risk and vulnerability especially when familiarity with the object is low. And trust leaves the future open. And confidence on the other hand means that there's high familiarity with the object, which can be about everything. And since things work in a predictable way, it causes a closed future, I understand it as an expected future. Anyway, trust and confidence, both of them depend on familiarity. And Patrick's skepticism here is that nobody talks about what familiarity is. So that's something to inquire about.
- Trustors who trust in systems change how they reflect on the system, how they receive it. And trust relations and expectations change over time, they are not predetermined. So that's the summary. System trust has a dualistic nature, the six major components of trust, we need to study expectations, and system trust relies on relations that change over time. And the conclusion is to advance the discussion on system trust and to distinguish it from other forms of trust, such as confidence and familiarity.
- And I was left with questions and it was really thought provoking. And let me just repeat the two topics, I have a few that I think would be interesting for us. So the first is about the approach presented by Luhmann of systems detached from human influence. And I think it's relevant to us specifically thinking about automation and smart contracts. And the second is these six characters of trust. And you think though, might refer to something more structured where there's some kind of a new phenomenon of trust that we did not have before, such as peer to peer trust.

Patrick Sumpf

- It wasn't only me in a way who said that this idea of genuine system trust might be worthwhile exploring more. In fact, Luhmann himself did not use the word confidence until 1988; his original paper in 1968 did not distinguish and he only used system trust and trust in systems.
- Is it possible to assess the output of systems? Do people perceive it that way? They do at least for the energy system. You brought up system definition and architecture. And this idea of a systematic interplay of elements is what I came to. Coleman talked about emergent systems, something that stands out and builds an emergent identity that is distinguishable from its environment. A system can be this sort of receptor of trust. It doesn't need to be a social system exclusively. It probably needs social elements, that will be the difference also to Luhmann who only talked about social systems mainly.

- And trust is inherently social, which is also very important. But this leads to its own discussion of sociotechnical systems, right? Like what elements are connected on what level, do they have direct interaction of elements or not? Or are they just what I call pseudo communicative elements when they're technological, for example, like, you know, some conservative scholars say you can't trust in technology because technology is not social. And then I actually said, well, it depends on if it can communicate or is perceived as that, because trust is essentially based on communication, especially in the system theory view, and everything we can communicate and can receive and send a message for example, that people perceive it that way. That was my argument that it can be at least assumed to be an object of trust by trustors. So I very much made this point that trustors create their own illusion, illusion of trust, because trust always has this kind of fictitious element in it. You know, you can't really say this is real trust or this is fake trust. And this is this whole idea of attributions, of what Primavera mentioned in the beginning, which I find very, very important, very underestimated in the debate when people say, Oh, this is a situation for trust, and this is a situation of confidence, and this is a situation of familiarity. I mean, you can do that as a scholar, you know, expost, but people do in reality, that's a matter of their own perception and their own attributions and how they deal with the facts at a certain time. And that structures the trust relationship, that's the overall idea. And these relationships last over longer times and years. My relation with politics, my relationship with the economy, my relationship with, I don't know, a blockchain system. It's hard to describe this stuff with the objectivistic terminology of scholars telling us this is the nature of trust or confidence or something. I think that's also something that's decided on the ground by trustees.

Primavera De Filippi

- I've been brainwashed by Luhmann and it seems to me that even though you can trust a system, it feels like the elements required for trust, it’s not just lack of knowledge and positive expectations for the future, I think there is one additional element, which I have a hard time seeing in a technical system, which is the possibility of actually being betrayed. One of the distinctions between trust and confidence is this element, which Luhmann calls vulnerability, which is more than just not having certainty, but by the act of putting myself in a situation of trust, I know that I’m actually giving that entity (system, organization, or person) the possibility to betray my trust. And that’s an additional step from just being uncertain. I can have low confidence in a system, but that’s it. Probabilistically I might be willing to take that risk, but if I don’t feel that the system as a system could betray me, and the only thing I perceive is that the system might not work as expected, it seems that this element of vulnerability and betrayal is an important element of trust, which perhaps would have motivated Luhmann to claim that when we are talking about actual systems as opposed to social organizations, it makes more sense to talk about confidence.

Patrick Sumpf

- I might call it disappointment. Betrayal is more normative and personal; a person can betray you but a system cannot. It's not possible to be betrayed if you trust a system. You build up expectations and then eventually become disappointed, but at this point you find 'access points' or people/organizations who embody a system if you want to have contact with a system. As a trustor you have an emotional contact; this is what happens when you engage with bureaucratic representatives of a system when they feel betrayed or the service should have been better.
- Betrayal is not always consciously reflected upon. They may slide into a situation of vulnerability (betrayal often only happens at a later stage), many trust relationships happen without noticing. But yes this normative component is at odds with Luhmann's and systems theory view. For him trust fulfills functions and solves problems. There's a contingency where you don't know how your opposite is going to react and what they're going to do. They have freedom of action and that's why you don't know what's happening. So you have to place trust at some points. It's kind of a more functional approach to something that isn't necessarily needing to happen. If you want to build up complex chains of action and have people act on your behalf and have a system run and work and all that stuff. So it's a bit of a different avenue. I would say than putting this idea of betrayal or the reaction, for a trustor to betrayal, to put that center. That would be maybe more of a psychological question in the view of the systems theory approach.

Primavera De Filippi

- I think you are right, and betrayal might be too strong. It's not the sense of betrayal, it's the sense of the possibility of being betrayed, which is the sense of vulnerability. So I'm not saying trust entails the feeling of being betrayed, but the acknowledgement that there is a possibility that the system will betray me. And to me if I don't manage to assign an actual agency and intentionality, then it's just my confidence in something that failed so I did a bad assessment. But I was not in a vulnerable position. Beyond the act of broken expectations, the distinction is in the perception of the possibility of betrayal (vulnerability) by engaging in this active trust relationship.

Patrick Sumpf

- Yes the risk perception: what's at stake and what can I lose? It could be a personal feeling of betrayal, that's of course one legitimate thing that can happen, but it could also be like in many of the blocks and example it's probably could be that is my data at stake, right. Or can people follow up on my, you know payment information? Will I lose money? You know, like anything negative that can happen to me. And that's also what I defined for the energy system. When I looked at that, you know, the risk perceptions were, can I influence the system to perform worse? Like, do I have bottlenecks? Can I still get

electricity? Is radioactive waste taken care of, do I contribute to that with my behavior? You know, it could be anything, it could be, it could be different risks that people have in mind. And then that's linked to making a decision as soon as you have risks on your radar and you see, Oh, if I do this, this could happen. Or if I don't do it, then something else could happen. That could be bad for me. So it's, it's, it's a little more general, I guess than making it emotional or personal right away, without excluding the possibility of, of making it personal for some people. But I guess for me, it's more of a personal reaction to a failed expectation, some people might take it much more calmly or might not feel betrayed and others might feel betrayed, but yeah, I wouldn't generalize that. I think I would put it more broadly, a bit more broadly than that.

Eric Alston

- This discussion is actually to me, like really cutting centrally at the true distinction between confidence or an expectation surrounding a cold risk calculation about an outcome you want and an outcome you don't want, often tied to calculable benefits. And so this disappointment, this vulnerability for me is like this essential line between a strict risk calculation and something in which you're actually imbuing trust, whether or not it reaches the level of betrayal in every instance. But I think it's interesting that given how close bedfellows that confidence and trust are, we often get it wrong. I ended the last call ranting at Ori and Primavera about the fact that I trust my car, and I'm not sure I get that right, after having had this particular discussion just now. We often tend to imbue things, we anthropomorphize a system when we are engaged in the constant calculation of doing so.
- Two very quick examples that kind of emphasize the shift. Imagine we want to gamble. We all agree to put in the exact amount of money and a random number generator that's mapped individual numbers that we each have. We'll just spit out a number. It's hard to say that you trust in a human sense, that you're disappointed when your number doesn't click around. But what I'm describing is the integral process that underlies gambling systems that eventually people come to imbue with a high level of trust. If you talk to professional gamblers, the way that some feel about the cards, the way that the specific probabilities are realized in a given instance, it's imbued with all kinds of emotion. And so for me, it is operating in the same way that we're describing with vulnerability and disappointment in that it's triggering truly personal responses. But that might be irrational. It might not actually follow that it isn't an appropriate recipient of trust, even if the person is indeed engaging in the process of trust, because of this process of the anthropomorphization of the system processes of risk that are just pure numeric calculation of probabilities in my stylized example.
- And so I think that there's something really essential in Primavera's point about this vulnerability and disappointment that you've both been emphasizing. So I guess I'm just co-signing with the additional example of, we tend to anthropomorphize systems that

otherwise should be pure receptacles of confidence. They're getting some trust in there too, because of this sort of repetitive familiarization with the system. And gamblers came to mind when I was thinking of an example of people putting an emotional heft on a purely numerical outcome, setting aside the dice being stacked, the cards being stabbed, all of those things. There's still often an element of that, I like to think in it.

- My very minor clarifying question was, perhaps for people more familiar with Luhmann, this definition might've been screamingly obvious, but I was wondering if you could speak a little bit more about values in particular, in your trust architecture, as you define it, just how you see them and the role they play. Because I didn't see clearly at least defined for me and in your table, it's clear that values have a different entire kind of dimensional vector than all other three elements of your ontology. It's like values are doing something very different, at least as I understand it from your diagrams. So I would welcome your thoughts on the role values play in your trust architecture a bit more explicitly, because that was one area where I just wanted more, but fascinating stuff.

Patrick Sumpf

- I think the general separation is important between general and specific functioning. And I think that's what you alluded to, the value level is on the general functioning side. And this is what they meant in those classic texts. When they talk about confidence, what Luhmann meant with system trusts. They trust in the functioning of a system like the economy, the economy runs, you know, you can go to the bank, you can put your money in the account. You can trust that they'd send it, you get returned and all that stuff. And that's like more of an expectation in the back of your head. But it's just happening. And you know, so, but there's those sides to that. And the more concrete stuff then relates more to people and to roles that can also represent a system or to programs that kind of condition a system and how it operates. And then I said, okay, the value level, which is the most abstract way of forming expectations toward, because the idea is that something like let's say sustainability would be a value, right. Would be something on the value level. And then the thing is that a concrete action can often be declared to be sustainable by some people, but not sustainable by other people because sustainability is so general. Or if you talk about something to be just, for some people, certain actions are just. For other people their unjust and so, so if you think that a system represents a certain value on a very, very abstract level then that means, like, for example, if you trust that the political system is not corrupt, which I'm not sure how many people do, but just saying you know, then, then that will be a very, very general trust located at a very high level of expectation that can barely be altered, I guess, by concrete experiences that they make on the ground or would take a long time until that would actually change because it's so general you know, whereas you would, if you trusted a complete person on the other hand, which is more specific if they betrayed you, as we mentioned earlier then that would maybe shift pretty quickly into, into distrust or lack of trust for that person. So I

think that the higher you go on this expectation ladder, which is like a sociology concept. The more robust the general trust actually is.

- Security of supply, for example, is a value in the energy system. People trust in the security of supply, the systems safe and will always supply energy and electricity. That's a very high, it's also a very unspecific thing, right. They couldn't really say why or how. And then some people were able to, and then it gets more specific. And then I guess if you have both sides complete, then people can be said to really trust the system actively. That was my conclusion in the end.

Eric Alston

- This is a really interesting parallel between the role of courts in the United States in ensuring the integrity of system level outcomes as relates to the doctrine of due process, guaranteeing legitimacy of the system overall to individual participants. In the sense that procedural due process is a check in this specific functioning where all of the rules and procedures followed that this person was due in this particular outcome. In contrast substantive due process is viewed as a general check to ensure these value level outcomes for the system like justice, like fairness, and both the courts, the higher level checks on system level processes. The courts of review are doing this simultaneously, but I see a very interesting analogy in the sense that this is descriptive of all systems, but a court of review has actually developed this procedurally in the United States for a long time. This distinction between what's called substantive due process to ensure values level outcomes, and procedural due process to ensure all of these specific executable outcomes within the system.

Patrick Sumpf

- People have concrete expectations and ideas and associations with a lot of things surrounding the energy system, for example. I think that we should be more open to understanding relationships between something that could be identified as a trust object (which of course we really need to define and identify first, we can't just say it's there. We have to kind of prove it through empirical research) that it is plausible that people make this connection. And I think to me, that's all that counts. Do trustors make that connection? Do they identify that as a trust object and then yes. What was said in the comment by Brett that it's designers, owners, users of systems that are the recipients of the trust. That is definitely true. But what I found is that some of these elements are only representative of the actual overall system. So people use these as intermediaries. They think of designer designers. They look at you know, supervision, supervisory agencies. They look at the government, they look at their local electricity provider, but often in many cases, as a representative of something greater, and I was able to show in the last chapter of my work, that this, what they call something, what I call something greater, which would be the energy system is, is what people plausibly described to me because they talked about it. You know, this is only like one little piece of an overall situation and

they all play together. And, you know, they gave me these like hints where I noticed, okay, this is like something emergent they're talking about. And they are all aware that there's like things working together to make this happen. This output for them. And then this is actually what they do with their expectation.

Brett Frischmann

- First of all, you've sold beyond reading the rest of the book. So I only read the one chapter that they shared, but I want to read the rest so I will get it. Cause it sounds like a lot of the other chapters you're referencing are relevant to research and things I'm thinking about a lot, but here's the thing.Have you thought about, or did you get at how those feelings? So when it's subjective and it's personal, what the trustor thinks and believes and perceives like components of their will and how they kind of approach the notion of trust. Have you thought about how those things are engineered, manipulated, nudged? In other words, those become vectors for manipulation to create trust where it isn't warranted. So I understand about the energy system in and of itself. There are good reasons to think that the electric company and all the regulators who regulate the electric industry are limiting the kinds of manipulation or nudging or things that I call techno social engineering that you can think that are directed along those vectors. But, we're generally thinking about what trust is and trust in systems and how it works. Like I think about trust in my iPhone, right? And the iPhone represents the mediating tech that connects me to a whole bunch of different systems. Do I trust my iPhone? Do I trust the system that asks iPhone users whether they trust their iPhone? Sure 99% do. Whether they should, whether the trust they have is when they truly understand? Whether the network of third parties hiding behind the other side of the multi-sided market network that the platform represents, whether you ask them, do you trust the thousand trackers and advertisers and app developers that are connected to you through your phone, then all of a sudden it gets a much more complicated question.

Patrick Sumpf

- You can never bring complete order to this chaos, right? I've tried it with the architecture to supply addresses. And these addresses are what people refer to to reduce their uncertainties. And, you know, in the moment of trusting, that's what he calls his leap of faith. And that's what, when it goes down and it's like an intersection, what many trust scholars say. You need to reduce this uncertainty to be able to trust, to be able to act. And this is like the sociological take on it that I share with Luhmann and others. It is like trust as a function. It's not only a feeling. I mean, you can also discuss whether it is a feeling in general. And I know that this stance is popular and I understand it and I feel it too. And I know what it means, but I think from a sociology point of view, trust is important because it enacts things. It sets things in motion. It makes people act and do something that they wouldn't otherwise do if they didn't trust. And then these chains of action that

occur and they are set in motion. For example, when you participate in a system, when you go to the election, when you use electricity, when you use your phone when you don't care about the trackers and you just go on all these websites and you do all these things. That's the function for sociology that makes a difference in the world. And that's what we're mainly looking at. And so the question is how do people make that leap? What makes them, what makes them gain their trust? And I think that's what I tried to present with the architecture and those elements you see in it, that people refer to something like that. Like they say, Oh, I can trust, you know, that the energy supply is safe because there's this institution that guarantees it, or because my neighbor he's so smart, I don't know much about it, but my neighbor told me there's so many engineers working on it and they've got it figured out because they're so smart. So you trust in the role of engineering. That's what you trust in, or you trust in the organization, which is the supervisor organization. So I didn't look into this, like a normative concept of like, is that, is that good or bad so much? Or is that they, do they offer themselves to, for betrayal, but like, how does it work? How does it, how does it go about, and then maybe from there, if you have this analysis complete and you have an architecture sketched out for a certain field. Maybe also for your field. If you look at blockchain technology in a certain area, you could say, Oh, these are the actors, these are the players. These are the elements in the architecture. Now I know what people refer to and how they reduce certainties. And now I can go about, and the next step to see, Oh, like, does that make sense? Or how much knowledge do they actually refer to doing that? Or how much non knowledge do they absorb? Like what do they all not know? And never look into it, right? Like the trackers and all that and leave it in the background but still trust and act on it. And that's, that can be a big problem. All these risks build up in the background and that, that falls at someone's feet at some point. Systemic trust can build up systemic risk.

Primavera De Filippi

- When I say I trust a blockchain for instance, if I were to actually use trust in this sense that I understand, it would mean I actually trust every single individual that's part of the system. Perhaps actually, if I were to think in those terms, maybe I wouldn't trust the system, but because I'm not thinking those terms and I'm just thinking in terms of like, my expectation in a very broad sense of like, well, it has worked until now, it seems to have some economic incentives, and therefore I have confidence that it is gonna run.
- And so you might have confidence in a system, even though you might not trust individually all the actors that are actually involved within that system, just because you're not thinking about that. And if I were to investigate them, double click on my iPhone and look at all the actors that are involved and that potentially could betray me, then maybe I wouldn't trust them. And then maybe I would have less confidence in the use of

my iPhone. But because I don't do so then my expectation is that it's going to work. And to me, we've pulled back into the field of confidence as opposed to trust.

Patrick Sumpf

- A very great observation and a great comment. Thank you very much for that, because that takes me to one of the final conclusions I made in the book which I find highly important. And that was the idea that trust does not mean that there can be no other forms next to it that referred to other objects of the same overall you know, framework like distrust or familiarity can also exist parallel to trust and toward other objects that are related to the same overall object. So basically in your architecture of trust that you sketch out and you would have to exactly assess and monitor and analyze what you just mentioned which is like this trust means that people trust a hundred percent. Is there no exceptions, is there certain circumstances that they trust under or is there certain things that they're distrustful of, other things that they're just familiar with and just act upon and don't really reflect it very much.
- So in the energy case, it was very obvious that people were, you know, distrustful toward the government taking care of the radioactive waste, for example. So even though many people were happy with or trustful with the security of supply or the idea that, you know, let's say green sources were used more and the electricity they expect it to be more CO2 neutral at the same time, they were distrustful that the rate of active waste problem was taken care of. it's a mix of different attitudes of the same phenomenon at the same time. And that's also something that's underestimated in the debate. I think I had some references from like Lumino, I believe and some other folks who've been bringing this up. The idea of separate yet related attitudes toward the same object. So, yeah, I very much agree with that.

Michael Heidt

- We talked a lot about the problems that you get when you use systems in that context. And I and I see the problems like basically it's once you go into Luhmann, like you have no subjects, you have no humans, so there are no, well, no, there are no decisions or motivations proper and so on and so on. So I was wondering. What are the positive contributions of systems theory, or why use Luhmann in the first place? And that it might be that when you go into sociology, And you were looking for it was a universal problematic of trust or the embarrassment of riches, really. So you kind of end up discussing Luhmann, but it's also with this notion of systems, I sometimes feel it's, it's, it's dangerous. Like it automatically creates confusion because if you talk about systems in the context of technology, it usually means something totally different. Like in engineering, for example, a system is a set of interrelated components, which together exhibit a characteristic that none of the individual components alone could produce or something like that. And sometimes in the text, the notion of system is closer to that than

it is to Luhmann. And in other parts, it's more like Luhmann. And, but for me, actually, one of the positive outcomes is that you get this, it yields this concept of complexity and complexity reduction. So I would be interested if that is something important to your approach as well. And also what kind of intellectual work, there's the concept of complexity Luhmann does, because I also just read chapter two. And you said you had a chapter later on, but the concept of systems in more detail, I didn't read that Obviously. So if I remember correctly, your concept of the system, it's based on emergence, but I wasn't clear if it's still about auto systems or not, which would imply a different conception of complexity, I guess. And further on in that vein, if you are talking about vulnerability, it might be interesting to look at the systems theoretic concept of coupling and resonance frequencies because it's another way that systems are interrelated and coupled. And also if you then go on and talk about technique technology, its systems Is it worthwhile to go back to the concept of auto systems because they, because everyone seems to be familiar with the new one, like Luhmann lifted his concept of systems from two biologists. Because systems could be something physical and for Luhmann, they could not.

- It would also be interesting to consider the distinction between complexity and complication, just to contrast or establish this distinction between yeah. Systems and, well, I don't know how to frame it right now. Confidence and so on.

Patrick Sumpf

- You made very fair points, but obviously to start off with, I mean, very, very general questions that have been discussed in decades of sociological debates, obviously which I also enjoyed doing, but I think it would a little bit you know, blow this conversation today. But what I can say is that going back was not my goal. I mean, I had this situation where I had systems theory and I had trust theory, right. And the trust discourse. And that was a little tough to bring together. I mean, if you want to really challenge the general notions of systems theory and stick to that, then of course it would have then made sense to bring in all these other scholars. But I tried to connect more to the trust community and to trust scholars. And I think for them, it was already probably you know, a lot of stuff that I brought in on systems theory, especially later chapters also, what you exactly correctly said, like the definition of Luhmann on systems and you know, what, what is the element of a system and what does it do? How is it perceived and what kind of system can be or what can it not be? And I discussed that and I tried to be pragmatic.
- It says for a system to be an object of trust, it needs to build a stable identity and emergent system. Like this is available to trust or that's a commonly shared background reality. So my idea here was that, you know, there can be systems of different nuances of different types of different couplings. As you also very greatly mentioned. I have a table here that I'm looking at the names of five criteria for system building tight and loose

coupling. As you said, symbolization of risk how's risks, symbolized, and that system functions and services. What functions and services does that system offer? Semantic advancement is important because it's a good indicator. How developed the language is in the, in the reality in society. It's not a coincidence that people speak of the financial system out there in the world. People speak of the energy system. They speak of the environmental system, but nobody speaks to the political system. Right. Nobody speaks up. Well, sometimes people might speak of the science system, but rarely. So I think that makes a difference. I think my indicator was the more systematic something like this interplay of elements and system can have, the more likely it is that that semantic is also available in the world. And so my definition is pragmatic and is based on this emergent idea. And I know that lumen would never say that, you know, elements of a system can directly touch, like for him, as you say, communication was communication, thoughts were thoughts and technologies technology. But I don't blow that distinction, but I'm only saying for the matter of trust and for a trust relationship, with a system that can be a system with mixed elements you know, if there is a systematic interplay between these elements. And that has made it available as something stable as an identity for trustees to refer to in their own world, in their trusting world, under these circumstances. That's what I sketched out in, in a, in a pretty good chapter. I believe I'm pretty proud of myself actually. What I thought was because I really struggled with that as you know, because of all the points that you mentioned, because it's a tough, tough one to go through and with a lot of history.

Matt Prewitt

- I had a few more thoughts on the idea about compulsory trust and it does seem like, I'm a little bit I'm a little puzzled by the emphasis on choice, in the context of trust, like it always has to depend on on a choice, because it seems to me that like a lot of the most important trust relationships, which are what I would consider to be sort of like genuine trust relationships in life, are not choices. Like we don't choose our parents. You know, when I, when I choose to, to trust like that the trackers are tracking me or whatever, but I, you know, it's, it's, I don't have an alternative, I have to exist. I have to use the internet. It's not fully compelling. I mean, there are choices in how I think about it, how I relate to it. There's a choice to sort of accept, you know a situation which I'm partially compelled into. But yeah, I I'm, I I'm just, I guess I'm just curious Patrick, about your thoughts on, on the importance of, of choice. Like why, why is choice important as like a prerequisite of, of, of trust?

Patrick Sumpf

- Yeah, thanks for that question, Matt. I think it's a very important aspect of mentioning and it, I think what I'm trying to answer, and that was also, for me, it was something that clarified a little bit my view on trust and system trust, which is again what Primavera

mentioned at the beginning of this that's attribution concept, I think because choices are not attributed to decision-makers or people who make choices and that happens if you want to, or not. For example if you, with the trackers, right, maybe you don't choose in that moment, but if you claim that, you know, you were betrayed for example, and you make that claim that, you know, something went wrong and they get your data or something, then it depends very much on the, on the societal circumstances. If that is the attribution that you make there that you say, I didn't have anything to do with it. If that claim is counted legitimate in society, and that will be a sociology question important for system trust. And I would say that at this point, the logic is very much that people could have known, people could have known that and they could have avoided the service or like what the, what some service providers do they like that you declare your agreement with the terms of service, right? That's another instrument there too, to construct a choice that you made. Even if you think you didn't make it, it's a constructive choice for you that is attributed to you either by the legal system or maybe you and your friends will say, Oh man, I knew about that. That was your fault, you know? And then that's what matters. And that's the sociology issue again, that's what matters, that attribution, how it's made, especially by your environment. And, and what's kind of legitimate and whatnot. And especially in the field of internet that you guys are in. If I understand correctly, it's a very, very delicate matter of how much responsibility do you grant an average user, right? I'm just publishing a paper with a friend in global perspectives, an Australian journal about this question of trust with online platforms. And then we have that situation where people are agreeing to all these things that are happening and then they will click it away on their screens. And then basically the providers are, you know, they're, they're not, not bound to anything and all the responsibility is put on the individual, but that's what matters, you know, that's what legally matters. And, and that's what brings the transaction into motion. And then what makes things happen. And it gets shifted in the background.

- So the choices are constructed and attributed. And that's very important and very underestimated. It's not that there's not a nature of a choice. You can't say, Oh, that's a choice. And that's not a choice because people will have different perspectives and there's different observers in society. And they make attributions and those make a difference for you.

Matt Prewitt

- I see the importance of distinction in terms of attribution of responsibility. But I think that there's another, there's another lens on it. Like, if we just want to understand how trust is being built, how you know, how it's propagating throughout society, Then that less hangs on that distinction because trust is constantly compelled, and like almost, like a huge amount of trust that exists in society in my opinion is compelled. And many of the strongest trust relationships are basically compelled. Including familial and parent

relationships. You know, like you, you don't choose your parents, but you, you, you have to trust them. Another example that comes to mind for me is let's say you're in like the military you're in like there's life and death kind of situations. Right? And there's someone else in your squad who is a complete screw up. You have no reason to trust them, but you have to, you must. Because your life depends on them doing their job properly and, and genuine trust does arise through that. And, and, and I think when I look around at the trust relationships that permeate society, I see a lot of this.

Primavera De Filippi

- I'm reluctant to accept the notion of compulsory trust. Because for me it can be divided into two different things, which do not feel necessarily like trust, and one is Matt's example. Actually, I wouldn't say that I trust my family because I'm actually just so familiar with it. And like, I've been living with them for so long. And I just, I just know how they operate. I know how they think. And, and, and of course there's an element of trust that I have built over time, but I managed to build this trust over time because I built familiarity with them. There is a coercive element initially, which enabled me to build familiarity which enabled me to build trust.
- And then there is another version which is what Patrick said which is actually coercive, like if I'm, if I'm in jail and you know, someone is bringing me food. I might not trust that person at all, but that's pretty much all I can do. And so I'm acting upon a situation in which I have no choice, but that would, then I wouldn't say, this suggestion does not actually lead to the construction of trust. And so I think compulsory situations need to generate familiarity to build trust. If not then it's just coercion. I wouldn't interpret an action that's done because that's the only thing I can do to be driven by a situation of trust.

Matt Prewitt

- But wouldn't you, would you say that a one week old baby trusts its mother?

Primavera De Filippi

- I wouldn't even know if trust is felt when you are one month old. I think it's, yeah, I don't know. I don't know what's happening in the minds of young babies. But, I think there is a familiarity that emerges before, and because of this familiarity then it's also possible to know that those people care for me, and I'm willing to put myself in a situation of vulnerability. A one month old kid is in a position of vulnerability, whether they want it or not. So they are coerced into being taken care of by the mom, which creates familiarity, which might become a seed for trust in a later time.

Patrick Sumpf

- Familiarity is a very underestimated concept in the debate. That's also one of my conclusions. My eventual conclusion is to basically abandon confidence. We talked about trust, specific in general, so it can be more abstract or more specific. Within the architecture where the abstract would be the confidence thing, the specific is the more active, decided trust in that regard. And then the alternative of the other phenomenon is familiarity indeed. That's also, that was my conclusion also in that debate. And I also use the example Primavera used myself with my family. And I said, I don't trust my mother. I know that my mother will not betray me. And that's a difference because if you know, then you don't reflect risks and you don't reflect the scenario that, that it could even go that way. And if you don't do that, then you don't trust because there's no, there's no alternative. There's no, there's no, there's no other scenario. There's only a good scenario. And if there's only the positive scenario, then, then trust is not the concept that you are looking at. And also not if there's no alternatives. As you mentioned earlier, Matt, when you said, Oh, what, like, what do I do right when this is just happening and I have to, and there's no way out and all that. That's also, I think something, some authors, classically have, have emphasized that, but then you're not in, and that's just what Luhmann said. Right. He said, if you're not in a situation of choice, you're, you're not trusting then you're, I don't know. You're hoping or you're, you're just, you're just doing something or you're, you're, you're, you're, you're compelled, right. So that concept, I believe, should be explored in more depth.

Ori Freiman

- So it's like suggesting that the concept of genuine trust, which and had, I'll give you a compliment. Your text was cryptic to me, like really, really hard for me to understand. I had to read it really, really slow as they do in big authors, like Foucault. It's not something where you stop and you hesitate and why did they believe in that? So in that respect, it's like, you know, a lot of thought-provoking discussions here, we're worried about genuine trust being one of them. And the question of an inevitable trust. Genuine trust in analytic philosophy is at the heart of many, many debates and they think they can begin to have an effect on these discussions. And so the vision of trust, the concept of trust is, is recognized as some kind of a normative expectation from a person. So in this respect, the baby would not have a normative expectation because it wouldn't consider that. On the other hand, the concept of reliance represents some kind of expectation of a certain regularity to happen. So there's this interplay between a system with decision making and the system without, because it's reliance. And this distinction between trust and reliance is common and I'm sure you might know about this and stuff, but the thing is that the concept of trust traditionally cannot be directed at systems. It's like, this is the humans behind the machines that manage them. And the way to bypass it is to reduce the object of trust from systems and technologies to normative expectations for the people

and institutions behind them. And then we can trust, you know, technological systems, or, or even artifacts like Alexa.

Patrick Sumpf

- Reliance, I've been into that, but I don't really remember actually the, that the, the distinction at the core, but I think it's, as you said, it's in analytical philosophy also, right. Or use there or something, or was that the concept of normative expectations?

Ori Freiman

- The basic distinction that many of them make is that trust is having a normative expectation for one person. And reliance is some kind of expectation for a certain regularity to happen. You can expect that without normative expectation and it's like that.

Patrick Sumpf

- I guess it's a little bit like hope isn't it. You know, that you just think it'll happen or it's a fulfillment of the technicality, right? That's what you mean. It's a regularity or technicality. Yeah. I mean, I think, I think it depends if, if contingency is involved, I mean, this is the basic argument from, from, from Luhmann, is that contingency involved, in the sense of, is there freedom of action on my counterpart? And so reliance, I guess, is a very narrow corridor and does not really involve a lot of contingency. Right. A lot of surprise, a lot of things that can go in directions that you can't really foresee. And so that's probably, yeah, I mean, very simple processes, maybe it could be a bureaucratic norm, right. That a paper is handed in and then some process is, is, is initiated or something like that. I mean, that is, I think that's actually just interpreted into my framework. It would be the general functioning, right? It's the general abstract level of functioning. That's the reliance. I can rely on the bureaucracy that they will issue me my passport as a citizen, if I, you know, hand in this type of work and that will be at my place in six weeks. So that's very abstract. And the reaction to your disappointment, that will be different as if you had, I dunno, a much more concrete problem than that right. If they monitored the building of a home and you relied on them to, to really have an exchange with you and an interpretation of the law on, well, they permit you to build your house in a certain way for it, right? That's, that's going to be much more specific and concrete for you, that experience.

Ori Freiman

- This is where normative expectations come into play and change the game from mere reliance to trusting.

Patrick Sumpf

- Expectations are the basis of all trust, right? So it's also something that I've used, and in lumens world normative is, is actually distinguished from cognitive expectations. I didn't bring it in, but it's a famous distinction in systems theory where it says normative expectations are those you don't change when you are disappointed by them. So it's like law or legal things, like you expect to, you know, there are certain norms that are always there or you always have to follow, you always have to go across the street you know, on a green light. And if you do it on a red light, you're not going to go by the red light in the future from there on, but he was still going to go on the green light. Right. So that's the difference. Cognitive expectations are like learning. So if you, if it gets broken, you change your behavior and you do it differently from that day onward.
- It would be good to get the expectation concept, even more nuanced. Than what I've done, I've nuanced it in that I said there's expectations directed at four different levels of generalization, which is persons, roles, programs, values. And that, that helps us in finding the right objects to look at when you talk about abstract trust, because people didn't know what trust is directed at. So I tried to solve that problem, but normative vs cognitive expectation is probably another level that's worth exploring. So yeah, but I haven't given it much thought so far.
- This was great. I mean, it's a very important topic. I'm glad that you guys are tackling it and I hope it can be helpful for you. I mean, I think this blockchain discussion from what I understand requires innovative approaches when you talk about trust. And I think the socio-technical idea, it could be helpful. I'm not sure how far it's only technical or the social elements, even. I'm sure there's a whole framework around the blockchain.

June 10 — Simon (2010): "The entanglement of trust and knowledge on the Web"

Attendants: Michael Heidt, Wessel Reijers, Primavera De Filippi, Ori Freiman, Morshed Mannan, Nathan Schneider, Philemon Poux, Paula Berman, Eric Alston

Key concepts:

- Knowledge and trust are fundamentally entangled
    - My research for further support of a claim is in principle open-ended. At a certain point, I have to stop searching for further confirmation and start trusting.
- We make use of knowledge (e.g. sources of epistemic content) to assess epistemic claims and rationally place or withdraw trust.
    - While sampling, i.e. the assessment of the credibility of a claim by comparing related claims made by several sources, might work in other settings, this method

is often questionable for assessing content on the Web due to the common practice of copying and pasting content without giving reference.

- Epistemic vigilance: since trust is not certainty, it involves the risk of being let down. As responsible knowers, we have to be aware of this risk. We may have a default to trust, but also a duty to watch out for signs of dishonesty and incompetence.
    - Such epistemic assessments require transparency, especially when trust is placed in unknown human epistemic agents or in non-human agents, such as algorithms.
- This risk of misplacing trust exists on the Web as much as elsewhere. The Web is an enormous conglomeration of epistemic content of varying quality. If we pursue epistemic goals, it is therefore crucial to extract valuable content from the overabundance of existing content. In many cases we have to trust, because we cannot check everything for ourselves.
    - Trust in algorithms: different algorithms are used for various epistemic purposes. They are used to provide new informational input as well as to assess the trustworthiness of information and its providers.
    - People trust the content of Wikipedia because they trust the processes of Wikipedia. It is a form of procedural trust, not a trust in persons.
    - Instead of trusting the content on Wikipedia, we now have to trust the mechanism assessing its trustworthiness. That is, instead of trusting the mechanism by which information is created on Wikipedia, we now have to trust the algorithm that simplifies a multitude of complex editing patterns into a tripartite signal.
    - Decisions embedded into algorithms are even less subject to critical scrutiny than Wikipedia articles—because they are less visible.
- To be responsible knowers we must not only be willing to assess whether we are warranted in trusting epistemic content and its providers. We must also make our methods for assessing trustworthiness subject to scrutiny.
- The users are not the only ones who are responsible for their ways of trusting and knowing. Those who design systems are also to be held accountable. The development of tools that empower users by making functions transparent and providing choices should be an epistemological and ethical goal for designers.
- Indeed, the extent to which tools enable users to modulate the amount of trust they can put in content should serve as an additional criteria for judging the quality of Web tools.

Transcript

Michael Heidt

- This is about knowledge on the web and how phenomena of knowledge and trust interrelate. So as the basic example, there's a statement on Wikipedia. How do I know this is true? Or what does it mean for me to know that this is true, and well, as we all know, not all of these statements are actually true, just as not all of the articles in the

Lancet or any other academic journal can be said to be true. At the same time, in order to participate in scientific discourse, I have to rely on these platforms and rely on a huge number of claims and agents whose reliability I myself cannot verify.

- So what does it mean to know? The paper first introduces what it calls classical ways to learn knowledge, namely perception, memory, inference. So we have these ultra classical examples. How do I know it's raining? Um, I might look out of the window, see that it's raining, then I know it's raining, right now. One hour later, I might remember that it's wet. And then I would have knowledge that it had been raining, or I might've spent a few hours in a windowless room, come out, see that it's wet outside. And then, uh, infer that it had been raining.
- These classical sources of knowledge do not immediately solve our problem though. While perception, memory are still necessary, it is not possible to perceive, remember, or infer the body of knowledge present within society necessary to take part in a scientific discourse, for example. In order to order belief as knowledge, we have to trust the testimony of others, but what is it that we do when we receive knowledge on the web? In order to clear up this question, we look at the conception of social knowledge. "I conceive knowledge to be a social status that can be ascribed to epistemic content by a community. Knowledge in this sense is a success term labeling epistemic content that has survived critical scrutiny from multiple agents and satisfies communal standards." So the concept of knowledge has to be examined in a relationship to a concrete community. Knowledge then becomes a specific type of content that has passed these communally accepted tests. This fits nicely with the Wikipedia model in which the community of writers and editors continuously enforce shared standards. Wikipedia content consists of surviving articles, those who have stood the test of time enforced by the community.
- So how does trust come into this? In an online community setting, we typically do not know who was making a certain claim. Yet we do trust claims made on Wikipedia or elsewhere on the web. In order to explain this, uh, the text mobilizes the concept of algorithmic authority to ultimately explicate the notion of procedural trust, algorithmic authority refers to trust put into algorithmic entities, such as Google, recommendations created on Netflix and so on. People trust in algorithms and platforms, if they perceive others to trust them. So again, we are faced with a systemic development here, for that perception to occur, others must already trust the system or at least behave as if they did so. So in the case of Wikipedia, what we trust is not an algorithm. It is also not necessarily the community itself. We trust the specific process, carried out by the community on the platform, the system in which a large number of editors constantly check articles and correct errors leaving only those intact, which adhere to communal guidelines.
- At this point regarding procedural trust, the distinction between trust and reliance does no longer apply since this distinction depends on the intentions of the trusted agent. Since processes and algorithms do not have intentions, this distinction collapses, trust and reliance become interchangeable, and the text refers to trust. It talks about both, this one

overarching concept. So to summarize, in order to know, we always have to trust an agent's processes and artifacts. At the same time, we use knowledge in order to place and withdraw trust.

- So how would we deal with this mutual dependency of trust and knowledge on the web? In order to address this entanglement in a pragmatic way, we then have to behave like good experimental psychologists and try to operationalize trust in a manner conducive to web platforms. So trustworthiness cannot be observed directly, but it might be possible to construct proxies of trustworthiness. The text goes on to provide a couple of examples. The first of which again deals with trust on Wikipedia.
- The first one Wikiscanner is simply tracking IP addresses in order to identify self-interested Wikipedia edits, thereby partly de-anonymising participants. The second Wiki Dashboard is a visualization device intended to allow quick analysis of editing patterns. This will allow users to quickly get an impression of the distribution of edits of a certain user or article. In the case of the dashboard. It is still mainly the user who interprets this visual data, translating it into decisions regarding trust. However, it would also be possible to go a step further, give more authority to a process in the form of an algorithm and to implement something such as traffic lights of trustworthiness. Simply map these patterns to red, yellow and green.
- The final example concerns collaborative filtering algorithms. As used in book or movie recommendation engines, these systems usually operate by looking at your past actions, computer similarity metrics, and then suggesting items that fit these past parents. However, it's also possible not to weigh items, but to weigh users, and rate them according to trust or trust statements. As a side effect, this can also put users situated to be highly trustworthy or not too trustworthy at all. Interestingly in practice there often are quite a few users which receive ratings at both ends of the spectrum. Users which are controversial. There are two algorithmic ways of dealing with those problematic, either through local or global trust metrics, the global trust metric assigns a value of trustworthiness for each user. And the local trust metric is on a per user basis. In the sense of how trustworthy is this person to me? Creating a binary relation. Different processes can yield very different results. There might be users who would score very high on my personal trustworthiness scale and still receive a fairly low global score. And thus not factoring into global recommendations. So what kind of algorithm should we choose? Simon proposes turning the decision over to the user. She presents an UI prototype that would allow users to choose between the local and global approach. The intention here is not only to present an additional choice, but also to highlight the implicit authority of the algorithm. Eventually it could also be possible to arrive at a more qualified form of trust into the algorithmic procedures underlying the platform.
- One thing that was a bit difficult, or what I was wondering about, was the problematic of language. Although the paper deals with entanglement and quotes a lot of systems theory, the language itself, at least to me, seems to be very, um, analytical and didn't really reflect

this well, the systems heritage, which for me, well, it could, it could be a problem. Like if we contrast it with the presentation we had, uh, last week, um, like for system theoretic way of thinking, might fit better to this problematic of entanglement, at least for me.

- More constructive, these, uh, UI prototypes or visualization prototypes, of course we could. Uh, well, think and try to apply these things to the, to the blockchain. A lot of people have, um, I already have, and try to create trust metrics for the blockchain or visualization devices for the blockchain. Tracking IP addresses usually doesn't work, but all, all these solutions could be transported to the blockchain.
- Other points are preaching to the choir. Epistemic, uh, content doesn't need to be propositional and true or false. Well, I have more of a critical theory, uh, Frankfurt school background. So talking about the truth content of something other than a proposition, um, uh, it's a very valid idea from like talking about false consciousness, or the truth content of specific social institutions and so on. I'm not Hegelian but this more holistic or at least anti reductionist approach would lend itself to that kind of the dialectical language itself. Like what you're, you're you'd have systems, uh, theoretical language or dialectical are, um, at odds, of course, but to me more accurately reflect this systematic approach.

Wessel Reijers

- My impression of the paper was that it, uh, that it had good parts and also not so good parts. Uh, so it was a bit, um, indeed I agree that it's kind of, it feels a bit out of line with the papers that we read so far, because it also doesn't really reference the literature that we are familiar with. But what I find interesting actually is that it does also touch upon something that we have barely discussed, I think. And it's maybe also due to the nature of the systems that we are discussing usually, because, uh, if I might make a distinction between what you might call like forward-looking trust and backward looking trust, I think so far, we've almost always discussed forward looking trust, which is about, um, trying to anticipate what will happen in the future. I think most of the papers we've read so far, and that's also what system theory usually is about. I don't think it's about, um, I don't think it is that much concerned with, you know, truth claims about past knowledge is more about predicting future events or making probabilistic, uh, calculations with regards to, to future events. Um, and that's, I think most of what we discussed in the past sessions, whereas this one is really about something that's already happened. You know, some, some event has already happened in the past. And, and, um, how do we trust history? Basically, that's the question that's being asked in this paper, I think, which is kind of a very different question than the questions that we have asked before and in a way I think the reason for it maybe is there's also because in a way, blockchain technology closes off the problematic of past, uh, past related trust so to say because, um, the whole kind of consensus mechanism is focused on taking away any ambiguity with regards to the events that are recorded on the blockchain. So there's no, there's no, uh, reasonable doubts in a way about that. At this point in time, uh, this amount of, uh,

Bitcoin was transferred from this address to that address, for example. So in a way this whole problematic that we have about history, uh, is kind of lost, um, in, in blockchain technology, but at the same time, it also imposes a very specific episodic structure of blockchain technology that is maybe the interesting thing to, uh, to think about.

- You could say present looking, I think, but it's hard to capture, I would say. Once nodes validate the transaction, the transaction has already happened, you know, so in a sense it's like, uh, it's like validating some action, some events, uh, that, that, that at that point lies in the past.

Primavera De Filippi

- I like this idea of forward looking versus backward looking. At the same time, I'm wondering whether, um, when you're talking about backward-looking, um, are you talking about something else? When you're talking about trust, you rely on information from the past in order to make a decision about the future. And if that decision is, I want to accept this as knowledge and then act upon that knowledge, it's still something about the future. It is an open question. I don't know if I have a opinion myself, but I feel that my intuition at the moment is more that, when I'm talking about backward looking trust, do I have confidence in whatever has happened in the past, in order for enabling me to have trust for whatever things I want to do, relying on this thing from the past.
- So backward looking trust is closer to confidence than trust, because if we agree that trust is something that needs to be betrayed. Judith actually specifically uses a broader definition of trust in which she says it's about betrayal and deception, being disappointed. So she's actually talking about trust in the same way in which we talk about confidence *and* trust, where confidence is leading to disappointment, whereas trust is leading to betrayal. And so I think at least my interpretation is that whenever she's talking about backward looking trust, she's actually talking about confidence, and then trust is only when you act upon that confidence in a way that you could be betrayed.

Wessel Reijers

- I see the point, um, but I'm not entirely sure whether I agree with it because it would assign a too pragmatic or functionalist value to historical knowledge, because basically what you're saying is that the, you know, our interest in trusting the the past lies in our interest in anticipating the future, basically, you know, so like there's a kind of functional value in historical knowledge and say like, you know, like we want to use this historical knowledge in order to, to do something in the future, but I'm not sure when it is actually the case. I think that, uh, because there's also like this, this big conundrum about what, whether we can learn from history in this kind of very direct sentence, you know, or whether, um, or whether there is another way in which we value the truthfulness of historical knowledge, there's not this functional equivalence with regards to, like, we need to know whether it's true in order to, to anticipate the future. I think that that's

maybe the crux of the thing, you know, it makes maybe the, the, the value of history a bit too pragmatic in a sense.

Primavera De Filippi

- Yeah, I wouldn't go as far as saying that the only reason we care about knowledge of the past is to predict the future, to anticipate or foresee. But I do think if you consider the acceptance of something, the assimilation of knowledge, as something that you shape all the time, your behavior about the future, then to some extent, it is. Do I decide to rely on a particular information and to assimilate it, even historical information, it's still gonna somehow shape what I decide to take as knowledge, and that's an action that is a future direction. Because then I would say something and you would, it'd be like, you're completely wrong. It's gonna, it's gonna affect my future somehow. In a very short time period, do I decide, do I take the risk? And therefore, do I trust that actor that is giving me knowledge to actually assimilate it as my own knowledge, or do I maintain a critical stance to it? It's like short-term actions. It's not predictions. It's literally, it's almost like present, like, do I do it? Or do I not do it? And potentially I could be betrayed, if you lied to me, you betrayed me. Uh, if you give me the wrong information willingly, you're betraying me. Uh, and, and I can choose whether to accept the information or not. Then that could be a leap of faith if I decide to trust you. Um, and then in a longer term, it's like the fact that you assimilated the knowledge of course also has repercussions afterwards on pretty much your model of the world, your conceptualization of the world, which I assume will affect pretty much every action.

Ori Freiman

- So first of all, the context of this paper, I think it's good to point out that it's written in 2010. Uh, so back when we can just start with something and all of those things, and the thing is it was a communal production of knowledge, and that's something that analytic philosophy had very hard times dealing with as she pointed out because of the context of trust and only relies on human, and we can betray. So the context of your production of knowledge is something that we can compare to blockchain. And, but that's not the only thing that we can, uh, compare you and. She had this place over the beginning where she puts the theoretical background of what she would like to do. And that she mentioned critical contextual uh framework, a feminist epistemology, one of the three pillars of feminist epistemology. And what's unique in that it gives the walls for the community to decide what can be counted as knowledge and whatnot. And in this, eh, eh, contextual emphasis, then those roles really care about how criticism is meant. So that the knowledge that we produce, that can be like a very good analog to what goes on in blockchain, you can go further into that. Now as Wessel spoke about forward-looking trust and backward looking trust, which I really like this distinction. Trust is directed at future events and it deals with expectations. As we saw in other sessions, I haven't been

here for the past discussions, but trust is discussed in terms of expectation. And it's not in principle having to do with things in the past. And reliability is discussed as, uh, with your language backward looking it's like this low, like that we get accustomed to because it happens. And in the notion of confidence and that's when things happen all the time, it's an indicator for the reliability that can be predicted. So, uh, in that respect I wanted to jump in and say that I really like this distinction because it makes sense.

Primavera De Filippi

- And so in your, in your view, uh, this distinction you see it as like two facets of the same thing being trust, or do you see it as two facets being one confidence and one trust?

Ori Freiman

- Um, it's two different things, in my view. Uh, and in the paper, she said that she's using a broader notion of trust. So it's easy to read it from the perspective, but in a critical comment, she made our lives much easier by saying that it's a wide notion of trust and it will encompass reliability as well.

Wessel Reijers

- I think it is true that the reliability is backward looking, but I'm just wondering, if you think about reliability, we base our expectations on past data basically, and also like in a kind of calculated way, but I was wondering, so, I mean, because like historians use different ways in order to assess the trustworthiness of their materials. The real interesting question is like, is there only one way to relate to this material or that other like epistemic ways in order to relate to this historical material? And that's because I was just reading 1984. This question is kind of dealt with, you know, because there's just like, Ministry of truth, which has kind of a huge Wikipedia community, but the point of this Wikipedia community as the ministry of truth is that instead of amending things for the better it's like the whole thing is kind of turned on its head, like amending everything for the worst, because it's like, it's everything from the get-go is like focused on creating falsehoods and so much that you know, that nobody anymore knows what is true and what is false. And the interesting thing is that it doesn't have to do that much actually with everyday life. It's not about like, what can I expect? Because they say about like how they, how they falsify all the statistics, you know, like there was this promise to be in the past and then they falsified and nobody really knows what's going on anymore, but it's not really the people that, that people base their expectations on the basis of either true or false knowledge. There is some other message in there which is about historical knowledge and also the way it's being produced and, and I'm not sure what it makes any sense, but I'm just trying to, to, to, to peer into like the possibility of there being all the ways in which, uh, we assess the trustworthiness of historical data.

Ori Freiman

- Yes. So you, you mentioned 1984 and those kinds of things that I think on the other scale, because she, she said, that's the thing about the communal production of knowledge, which it's an agreement that is, which process of critical deliberation and scrutiny that follows certain norms that, that she points to discussing about the, sometimes we fail in individually entries and stuff like that, but in general, those, those communal norms that, that discuss criticism and stuff like that, um, keeps the knowledge, uh, away from being a dictatorship that decide what's true or not.

Primavera De Filippi

- So if we try to, to understand this distinction, uh, backward looking, forward looking, confidence and trust, both of them are actually looking at past data, uh, but confidence relies on past experience. And trust of course also relates in big part from past experience or understanding of the past. They are both involving probability calculation and risk. The distinction between what I call confidence versus trust is that there is this, uh, agency that exists, which can provide betrayal. And so to me, when you're doing pure, backward looking things, if the thing is static, it's just past, it cannot, it doesn't have agency to betray me. So in that case, it's pure expectation of something, it's pure confidence and reliability, uh, because there is this lack of freedom of action in this past thing to act against the trust that I have given to it.

Wesssel Reijers

- The knowledge about the past does have some kind of agency. So like for example, the genocide in Armenia, you know, the way this knowledge is being produced, has some kind of agency on the people who, you know, uh, who, uh, who integrate it into their world vision, you know, and the, into the worldview. So if so, so for example, even though it's a positive event that doesn't like hurt you directly or something or affect you, like whether or not it is, for example, being recognized as true or false to some people has a huge impact, like more than anything that is in the present or the future.

Primavera De Filippi

- But this is agency not of the thing but of other people, so it is like does the community I live in recognize this as true or not. But the thing itself, the information that is provided, the article doesn't have agency. The past doesn't have agency in terms of betraying my trust. Uh, of course it can affect if people decide to assimilate it and accept it as knowledge it's going to affect me, but that is the people, that's the community, but you cannot assign this type of agency to the thing itself, the thing itself is now static. It has been recorded and it is past.

Wessel Reijers

- That's also the question of the past-ness of the past, of course, like this actual event that happened, of course, that that event has, that is not there anymore. That's true.

Primavera De Filippi

- It's the paper, the journal, the article that is talking about something. That cannot have an agency to betray me. If it does betray me, it already has done it by saying something false, but it doesn't have this power to act. It doesn't have any, any freedom. And the way you're describing it was not, you say trustworthy, and I think trustworthy is an interesting one because, um, I think when you talk about knowledge, it's not very trustworthy. I think when you talk about information and knowledge, it's actually, is it accurate or is it not accurate? And because trustworthy entails that there is agency whereas accurate is just passive. Is the thing accurate or not? And when we talk about processes and when we talk about algorithms, then there's a lot of discussion about trustworthiness of algorithms. And I think that makes a little bit more sense to talk about trustworthiness when you talk about algorithms, because there is somehow, at least the AI based ones, uh, they do have agency more so than just information. And so in that case, I could see potentially that the algorithm is betraying me because I felt it was gonna work in a particular manner. And in reality it's acting in a different way. And so it has the agency to betray me. And it's not a matter of like, is the algorithm accurate or not. That's a different problem. And so I think you can distinguish accuracy and trustworthiness in the same way as you can distinguish confidence and trust. And to me past knowledge can only be accurate or not accurate. Whereas future behavior or things that have the capacity of having a future behavior can be trusted and trustworthy or not.

Eric Alston

- This discussion has been useful because it kind of clarified for me what I found troubling about the definition of trust employed and it seemed to me like it collapses to one of two unsatisfactory kinds of definitions when I think about it enough. One is that she's simply describing the phenomenon of acting under uncertainty. Or differently. There's some line of knowledge or like a sufficient level of your accuracy or your confidence in that knowledge, I like confidence because it evokes confidence intervals. And so looking back at historical information, some historical information, we have huge confidence intervals around, such as very, very early dates in history. We know something happened. Think about geologic time, or think about civilizations who have very, very few records remaining. The confidence interval around the dates that we ascribed to those things that happened is huge in comparison to things that only happened a hundred years ago. And so I definitely liked that because as somebody who considers themselves as having almost a friendship with uncertainty. I really like uncertainty, unlike many, many, if not most people, action is sometimes painful because action crystallizes this shifting sea of beliefs about possibilities. At the time you act, you usually need to act with certainty. A rare

example is if I'm investing and I want to hedge an investment, I can go 80% in and hedge 20%. If those are truly my probabilistic assessments of the two outcomes that I'm investing in. Most actions are not like that. And indeed, if you act 80%, because you're only 80% confident that you're right in your information. In most instances, that actually deeply weakens outcomes, you want to act definitively even if your information is, you have lower confidence than certain confidence. And so I found her discussion to be, when you're acting under uncertainty. And I understand the vast majority, if not all actions to be undertaken in a condition of at least some uncertainty.

- So on the one hand, the definition collapses to simply, every time you act with uncertain information or information in which you have imperfect confidence, you are engaging in trust, but that's not a particularly satisfactory definition, because it seems to collapse with the definition of action in an uncertain world. The other possibility is, and this is what got me thinking harder. Are there any actions, if they're thrust upon you sufficiently suddenly, or if your confidence in that information underlying the action is sufficiently low, to where you're like, I have to act and I have no freaking idea what is the right action, but I have to act, but it's effectively a coin flip because my information, my confidence in that information about the relevant action is so low. Or action has been thrust upon me so suddenly, because there's an interesting temporal component to her definition of trust, which surrounds the ability to weigh all of these different sources of information to reach knowledge, to reach sufficient epistemic certainty to act, but is there a class of actions where your confidence in the information is sufficiently low, or you just have so little processing time, even if you could collect good information, you don't have the time to reach the relevant set of probabilistic assessments to give you the trust in your knowledge necessary to act. And so for me, I found her definition a little unsatisfactory because at least in my understanding, it collapses to one of those two things, either all action intrinsically involves trust, because we live in an uncertain world, or there's some weird line in terms of time or requisite certainty in your information past which you're engaging trust, and otherwise you're just a gambler or something? I'm not sure. So for me, those were my dissatisfactions with the definition as put forward.

Primavera De Filippi

- I felt she was claiming that you might act, or you might assimilate some knowledge, while knowing that, uh, it's not 100% verified. And if there is enough trust, you might do some actions, but not others. And you know it depends on the stake or something like, uh, if it's extremely important that this is true and otherwise it has a dramatic effect, most likely I will do more research and verification until, uh, until I feel sufficiently satisfied that I can trust. Whereas if it's just like a small thing that doesn't matter much, maybe I, I just don't care if I just find something on the internet. Okay. So I felt that she was not necessarily saying there's this one line and every action is based on the same line.

She might be saying that there is a line, but the line is variable depending on which action. And what are the consequences of the action.

Eric Alston

- Yes. And there's something in the chat that comes up closely related to this. Since the text deals with knowledge and relationship to science, does it even address the problematic of action' let's insert decision or action. So you make a decision not to act in Primavera's example, given just now, if you don't have sufficient confidence in the information, or maybe you act, if you do, and the lower the stakes. And I wasn't trying to argue that that line is always the same. I see that line is shifting fundamentally, depending, especially on the stakes. But all else being equal, I still see a problem, which is to say, is there any knowledge of relevance to this discussion that isn't a predicate to decision or action in some sense? Like, what are we generating knowledge for? If it isn't trying to inform our stock of knowledge to subsequently inform our actions or indeed, if we're more deliberate in the process and we're trying to actively seek knowledge in order to make a decision, that makes the point even clearer. I just wonder if there is a set of knowledge out there that people expend costly effort to become more certain of, that isn't related to decisions or actions on the part of that individual. I struggled to think of it, but I am the one with the possibly the most economics training on this call. So fair, like that critique is, is looming already. But I do wonder.

Primavera De Filippi

- I think I agree with you, I would just add the caveat that acquiring knowledge of the world is itself an action. I want to acquire knowledge because I want to acquire knowledge and that's my action. Uh, but then I will say that no matter what I want to do, the fact that I have a particular knowledge and a conceptualization of the world will nonetheless, regardless of my intentions, constantly affect my actions. With the backward looking, you're actually doing it because eventually it involves taking a forward looking step, which is the next decision or the next action. But I don't know.

Eric Alston

- I would say it's at least a fair point. If you read some of my historical works, many people would argue. This person seems to be interested in the pursuit of knowledge that has little to no normative purchase in today's world. And so I definitely intrinsically value the pursuit of knowledge, but knowledge that I'm solely accruing because of my own preferences. I like learning more. Does that implicate trust? For me, trust comes in when you have a stake, as you already emphasize Primavera of some kind as relating to a decision or indeed an affirmative action where those uncertain beliefs become crystallized into action.

Michael Heidt

- For me, it's kind of a different kind of trust that you are talking about. But I think something like that does exist in the scientific world this paper was talking about as well. Like I always conceived of trust as grounded in vulnerability. As she points out when discussing this problematic of authority, which I think the paper kind of problematically confounds, uh, algorithmic authority with trust in algorithms. If I, as a scientist, quote certain articles or based my research on some, something or refer to research my colleagues have done, the question is how vulnerable am I becoming by doing that? If I quote an article in nature, I don't have to trust the author at all. Um, I might even suspect that there's something fishy about the article, but it doesn't matter. I'm not vulnerable, if I quote nature, that's okay. If I quote an article on Wikipedia, it's not so great. If I quote a pre-print, that would maybe entail the necessity to maybe trust that it passed peer review and so on. So, um, for me, this problematic of trust then surfaces. And if this communal policy hasn't already taken my place, um, or if I am on this notion of risk, would that mean like trying to anticipate what this scientific, how this scientific community will behave? For me this notion of authority then eliminates trust. Authority increases the likelihood of being right and reducing the penalty of being wrong. And because it reduces this penalty, um, if I quote nature then I'm fine, basically, I don't have to trust anyone.
- She says that in order to increase the probability of trust, you need to have more visibility, more transparency. She proposes all those tools, uh, that enable us to better understand the process, to have more visibility on the actions that are being taken, so that you can make better informed decisions on whethe you wanted to trust or accept that knowledge. And, and so to me, I guess it is like creating transparency, uh, and creating like explainability or whatever, like more visualization of how things have come to be. It's this transparency about the process increasing my confidence and therefore, I will be more willing to make that leap of faith to do an action based on this knowledge that might be a trust-based action. Uh, but the process of investigating or the process of disclosing more information about the process by which the information has been created, for me, it's confidence building. It's actually not at all leaving any room. In fact, it's reducing the room for trust because it's increasing the confidence and therefore it makes it perhaps easier to trust because you need to trust less. Because the gap has been filled with confidence. Hence I need to trust just a little bit. There is obviously the question of trust in the people doing the process. And so you can have like a reputation system, uh, but in the end, this concentric system in which I need to trust the people that are participating in the process in order to have confidence in the process, in order to accept the outcome of the process as something that I want to rely upon in order to make a trust decision.

Morshed Mannan

- I find her arguments about, um, how, you know, trust in procedures implicitly means trust in algorithms. There's an argument that she makes that we are able to, um, see a rise in, you know, trust in algorithms because of the fact that for instance, with Wikipedia, we were able to trust in procedures. And I feel like this is something that moves too quickly because of the fact that partly as others have mentioned already, her definition of trust is quite broad, but I also think that it sort of, you know, I think it's a poor example to try to make this argument because of the fact that, um, I still think just as you mentioned right now, um, that it's just obscuring, um, you know, the interpersonal trust that actually existed.
- And one thing that I thought is interesting, it's one of the few papers we've discussed so far, which does talk about trust mediation and reputation systems. And I was looking at some of my notes from there, um, from his paper on trust mediation and especially looking at, you know, backward looking first and he says, um, something interesting, which I think I agree with more than this particular paper's articulation of trust in algorithms. So he says, um, you know, uh, platform marketplaces and resource sharing services act as institutional trust producers. Uh, they do so indirectly as technologies are used to filter how we present ourselves online to ourselves and to others. These digital filters and mediation are important to take into account when understanding interpersonal trust. Technologies, particularly the companies behind them, directly produce interpersonal trust. As the reputation systems are used by persons on both sides of a platform. Uh, but the tech company typically recedes into the background and leaves users with limited recourse. And then, you know, he uses this both to argue about the lack of agency of technology, which we've already talked about. But yeah, I think what I found to be particularly interesting is that he still uses this as a basis for, um, arguing that what we are trusting is not in the, you know, the algorithm itself as such, but either in persons or in certain institutions. And then he looks at certain tools and certain applications of those tools as well. And I think, um, given that both of them talk about these recommender systems and of course I acknowledge that they are 10 years apart. And so there's obviously an influence of that, that, um, I find to be, you know, more persuasive in trying to understand whether this paper helps us understand more about trust in technology.
- And second, like my last point with this that I thought was interesting, especially based on Wessel's point about, uh, 1984 was, uh, this, this very interesting, like, uh, science fiction short story by Carl Schmitt from the early 19 hundreds where, where a young Schmitt talks about this fictitious future human community called the, the Burri bunks who are basically scribes, who constantly just sort of archive their experiences. And all they do is, um, not really do anything. So this concept of action that we've just been talking about and, um, just, uh, they're busy with archiving their daily experience. And thus in a sense, I find it interesting because they have, I guess, a perfect knowledge, but that perfect knowledge, um, impedes, um, action rather than actually spurs future action.

And I think especially it has a very interesting discussion on sovereignty as well. It's written almost exactly a hundred years ago as well, which is really fascinating.

Nathan Schneider

- I mean, this is really just something I'm puzzling with, and really don't know how to conceptualize. But I tend to think that there is more than a difference of appearance between epistemic questions and the action question. And, you know, uh, as I said in the chat, one of the reasons is just an ongoing reflection on the relationship between this and class. That they're like distinct training, um, mechanisms that people in different social classes learn to evaluate reality and, and their, their relationship to action. And, um, and this really came up through a mentor who is a long-time trainer of activists and who just observed over the years, the way in which people with different class training relate to and emphasize action versus decision very differently. It does make me think that if we focus on trust in the kind of epistemic context, as opposed to an action context, um, even though, you know, it does seem like they are kind of reducible to each other, actions are based on how we think and what we believe about the world, right? Yet there's some, it does appear to me that there's something different about, uh, about a, um, operating with a bias toward action, as opposed to a bias toward a kind of abstract decision. And I don't know how to draw that distinction exactly. It is something that in the current topology I'm working on for the, uh, paper related to this project, I'm considering whether I can articulate a bit better in looking at some of the different projects out there that are focused more on decision, as opposed to action. And it may be that in looking at the designs that people are making around, um, uh, around these different platforms. So we may see some of the differences in what kind of trust we need to do this, as opposed to that? Um, you know, I, I mentioned in the chat also, for instance, that colony, which has a stronger bias toward action in the blockchain world, emphasizes reputation. Um, whereas Aragon, which focuses on decision, emphasizes stake. Those are two kinds of different theories of what makes somebody worth trusting. And, those correlate with different designs and, and, and intentions.

Ori Freiman

- And let me just quickly say that Aragon is facing a real crisis of trust.

Nathan Schneider

- Yeah. I just had a pretty awkward call with a group of the new folks.

Primavera De Filippi

- So you feel that, uh, action does require a different type or a stronger degree of trust than decision?

Nathan Schneider

- Well, I think that's something to explore, you know? Is there a reason why these two platforms that are looked at that have different biases, one production, one toward decision, use different trust mechanisms, right? Is this a pattern that we see replicated in other contexts? Are there patterns in action-oriented, you know, trust systems, um, and, and, and elective affinities also in more decision oriented, um, systems. I would hypothesize that you might see groupings and patterns form around, around each cluster, but, you know, I'm really not sure about that.

Eric Alston

- I think there's something quite interesting that this distinction has made me realize, and it brings me back to Michael's original motivating example of seeing rain occurring or seeing the streets wet, and the reputation you're describing relies on observable action on the part of community members. So is there a way in which knowledge or beliefs are different from empirical information in terms of the way they inform trust? And so one community provides an ongoing stock of empirical information that people can ably reference to generate reputation, or indeed the community does that in an automated way, so everyone can see someone's reputation score automatically. That relies on an empirical stock of information, but that seems structurally different from going out to Wikipedia and finding the answer to a question that almost by definition, you have little to no empirical information or experience with. And so there seems to be, to me, notwithstanding the fact that both are potential sources of evidence to give you a requisite level of confidence to act. It seems like they are different. And like the focus of the article seemed much more on the sort of knowledge stock and our confidence in the development of that knowledge stock and indeed who developed it. That being said, I do think that there's a bit of an artificial separation I'm laying out here. This dichotomy is falser than I might be making it sound simply because, I at least am in a camp that likes to believe we can derive knowledge from empirical information through reliable application of different techniques to generate that empirical information. And so it's, it's, it's interesting. I think I might be playing a little bit fast and loose by saying they're two distinct things, but the, the, the examples you gave Nathan made me think of that.

Primavera De Filippi

- If I can just interject here. I think that if I'm just seeing stuff, like experience, I don't need to trust or have confidence about anything except myself. I've seen something, I assimilate it, and okay, that's what happened. Uh, whereas if you tell me you saw something, then I need to trust you in order to assimilate whatever knowledge you're giving me. And then if it's not you, but it is like a very complex process of scientific verification and whatever. Then I need to either trust or have confidence in that system in order to assimilate that knowledge. So all of them eventually lead to me accepting or

not accepting the knowledge that there are multiple layers of trust and confidence in between as opposed to when I just experienced it myself on my own.

Eric Alston

- I think you're describing the at least half step, if not logical stutter step, that many ardent empiricists make when considering other disciplines as epistemologically inferior because they make the step of saying, well, yeah, if we all are independently observing all of our own data and only making our decisions based on that, that's a high level of empirical confidence. But the moment you get into a world where you're extracting data from statistical atlases compiled a hundred years ago, you're engaged in exactly the same process that everyone else is. You're just using a slightly different information source. And so I think that that's, I think that's at least interesting to someone who's more than a casual empiricist.

Wessel Reijers

- Yeah, just a wild speculation because I thought it's a very interesting comment you made Nathan, and I think you, you see kind of a similar distinction in philosophy in political philosophy, um, with regards to whether you have a proceduralist, uh, approach or whether you have more say a goods based approach approach. So, in this case you would wonder whether like Aragon is more proceduralist than, uh, Colony is more utilitarian perhaps, where it's, um, where in a way it's this distinction between the good and the right, I suppose in the background.

Primavera De Filippi

- To me, it is actually strange. It's a problem that people making decisions do not actually have the same standard of scrutiny and reputation as people making actions. Uh, first of all, because it's unfair that you're judging the actions of people, which perhaps came because of the bad decision of someone. And so I'm going to just negatively assess them or positively assess them but in the end, the decision that has triggered this action came from someone and, and maybe that person should actually be assessed, uh, as part of who contributed positively or negatively to the action. And, uh, and similarly, I think that actually why is it that people making decisions, uh, should be assessed by how much stake they want to invest in that decision, as opposed to how much wisdom they hold, and in that case, that would be like an accompanied reputation based system as well.

Nathan Schneider

- Yeah. And I would, I would even kind of reverse something Eric said earlier, which was about like decision needing to be really certain. I feel like this is a little embarrassing, but I, I found that distinction useful in the great bestselling book, which I've never read, thinking fast and slow by Daniel Kahneman, where he draws this distinction between fast

and slow types of thinking. And I wonder if that's actually what we're talking about here, where the fast is like, you need to act now, you need to do something, you are relying on like a different kind of knowledge and a different standard of evidence and conceptual coherence than you are when you're thinking slow and you're sorting everything out in your head. Wessel, when you were talking about, uh, you know, the question of like which political philosophy that says, it kind of felt to me, like something about that approach to the question I'm not sure about because any sort of political philosophy, in some sense, I think is in that slow thinking category. And I think what I'm getting at is something about a kind of a kind of pre philosophical thinking, that an action bias involves, one in which things are actually more probabilistic, um, in which we are relying on a lot of uncertainty. Um, but we are creating a kind of networked knowledge through it.

- Again, I think there's a class distinction here in that different kinds of class experiences often, uh, you know, the kind of thinking you can do sitting in an office is very different from the kind of thinking that you can do on a factory floor. I think I'm getting at some of that difference, where it's not just, you know, which structure of thinking are you doing or which philosophical school are you part of, but which form of, of epistemology are we, are we engaged in? And that, if you are, um, in, in a situation that demands action in an urgent fashion, um, there's a, just a whole other set of standards that, that apply in, in forming kind of assessments of trust than in, um, a more conceptual, longer thinking frame.
- And I guess to bring it back to the article, you know, as I was reading this, it felt like very, very biased toward that long thinking conceptual frame. Um, it didn't feel very relevant to the action frame. What do you do right now? How do you make quick judgments about what's going on in Wikipedia right now?

Eric Alston

- I think Nathan you're making a very important distinction, which is how long you have to make the decision before you act. But the point I was making is at the point you convert your epistemological level of certainty into action as a result of a decision. And your decision might be to act now, to not act at all, or to act later per my comment in the chat. But if your decision is an expectation to act nonetheless at the time you act, it's odd to think of your volitional force as being 80% when you act, if I choose to act carefully, I am acting 100% in line with my intentions underlying my actions. A careful action is not an 80% action. An 80% careful action is like taking some unnecessary risks in pursuit of your objective. But if you've decided you need to take careful action, you should act with 100% volition. And so it's that sense in which I was arguing that action is weird, but I agree with you, the question of how long you have to decide is a very important one, because that's one of the areas that made me come down with her definition of trust,

which was it's either that, or it's all decision making under uncertainty involves an element of trust. And I'm not sure I quite like either of those definitions.

Ori Freiman

- The decisions are not individualistic. That's, that's what's going on with science and Wikipedia. The atoms of it might be individual decisions or individual actions. I don't know how to differ between the two, but it's not only, it's not individualistic. It's highly political, both in science and in Wikipedia. The critical contextualism and empathy is democratic in its nature. It's famous among the circles of those things, but it's famous for holding four tenants, which are very democratic in the political manner of human interaction between us. So it has to be a public venue of criticism, members must respond to the grievances, so it cannot go astray, and there should be standards, which I'm known to as one, and members agree on those standards. And the fourth one. Which I think is the foundational one, even though it's fourth, is that there's no one intellectual authority that's alove all the others. And so it's like every actor, every agency, or every person participating in this game has some equal authority in regard to having decisions about the community production of knowledge.

Eric Alston

- But in terms of the article we read, I saw this as distinct. If this is a component of the process of knowledge production. Ori what I would I take you as describing and correct me if I'm wrong is a normatively preferable production process for knowledge like a knowledge process that has these characteristics is one in which you can trust more or one in which develops better epistemological results, um, epistemic results. I at a minimum am not in the camp that is sufficiently confident to choose epistemic versus epistemological, but that's obviously a function of my, uh, my lack of grounding in this, uh, in this area. But would you agree, are you describing the production process that gives greater trust or confidence as we've somewhat described already?

Ori Freiman

- Yeah. It's an intuitive answer. Those rules or those foundations do help us because, because we need to trust the say so of others, like, like the distinction with confidence that if I have my own experience, there's less, less place for trust. But in, in a community like the scientific community, we will have to rely on someone a few years ago about their testimony. So because of that, those rules help me trust the say so of somebody else because there was the community to supervise that, the outputs reached a certain threshold for me to adopt.

Eric Alston

- So let's use the example of the knowledge stock of a particular community as being passed through oral tradition. And this is a community that includes a lot of myths about why things are occurring, totally different community. They're modern, they like empirics. And so they rely on data. Other than the normative aspects of the process where it's like, we might prefer this process over here, or we might prefer this one, I'm not taking a stance on which knowledge production process is superior. Do both of those involve just confidence in a backward looking sense. As we've already described is one, trust is one, are they both trust? I, uh, if I'm in the camp, I'll, I'll, I'll plant my flag. I think they're both confidence. Although the one with the oral tradition to me, approaches trust more closely because of the way in which the tradition is being narrated by specific individuals in the community that might be representing that knowledge in ways that invoke this kind of personal relationship that we've described at length so far. But I'm good at generating examples that at least make me uncomfortable with my own definitions.

Ori Freiman

- Can I challenge you, it's not about data. That's the community that has data, that is not the criteria. And then think about a community of palm readers, which they have data and they agree, and they will do something, but it's not aiming for truth. They produce something that they will call knowledge, but the aim of it is not truth. At least not in a...

Eric Alston

- Would they say that? Would the people engaged in the production of that knowledge say, or I guess we're getting at a deeper question about whether or not they think they're charlatans, but like, I, I generally don't assume that people who appear to be in the good faith production of something all believe that they're deceiving their customer base. I also don't think a customer base can last that long if they all believe they're being deceived on some important margin or indeed are actually being deceived. Do you already think that all Palm readers believe that they are not engaged in the production of knowledge?

Ori Freiman

- Yeah, no. That's the thing, that if you have a community that believe in palm reading that it is true, they believe it, they produce knowledge. But the distinction between the scientific production of knowledge and other bodies of knowledge and religion included is exactly those shared standards that go through criticism, through public criticism. Um, I don't know if it convinces you.

Eric Alston

- No, no, I, I wasn't attacking the normative value of one set of the piston illogical production standards, especially around the area that I'm producing stuff in. So like, I am

not here to assail that particular pistol, the set of production standards by any stretch. No, no, no. I was saying it does it though. I just see it as a distinct inquiry, which is which production standards are normatively preferable versus does any set of production standards not involve trust on the part of the recipients of that production when they go to engage in action with the fruits of that epistemological production?

Primavera De Filippi

- This is a tricky question, because what you're asking me is actually just, uh, doesn't every action or decision involve trust independently. But then all the questions about the production of their epistemological knowledge doesn't matter as much. Even if I have knowledge that is only my personal experience, every action will still involve a particular level of trust because I'm still putting myself somehow all the time in a situation of vulnerability.
- I wanted to, to jump back on the thing Nathan was talking about, this thing of like decision versus action and at first I thought I wanted to say that, um, um, decisions actually enable you, like, if you don't need to act immediately and you, you get to have time to decide. It could lead you to less trust and more confidence because you can take your time to analyze more things. And so you'll actually reduce the degree of leap of faith that you need to do.
- At the same time, I actually think it is also the opposite. The more you spend time making sense of decisions, the more you realize that they have so many variables and uncertainties and contingencies, you know, it's like blockchain. Either I have confidence in blockchain, and I'm just going to use it because I don't know exactly how it works, but people say it works well, so great. But if I started investigating, and I'm like, oh my God, this is like, I need to trust all those people that I obviously don't trust. So in some way it is also the decision of people, because they start spending so much time digging and digging and digging. They actually have more of a need for trust as well because if I was, I'm just crossing the street, because I know that the green light is okay. But if I were to analyze, is the green light actually okay? And what about this and that, then I will have to trust so many things. So in some way, I don't know if there is a fair assumption that can be made, saying that having more time to make decisions reduces trust and increases confidence, uh, as opposed to vice versa. I don't think we can assume anything.

Eric Alston

- Yeah. Economists called us the paradox of choice. Like it presented with too many alternatives to satisfy your needs. It's actually destructive to your ability to truly satisfy your needs because you're hungry and you need to eat. And there are a thousand candy bars to choose from. Eventually you're going to collapse it to a few or, but the same point pertains to information sources and improving your certainty to action. Although Primavera, I see you as also making a distinct point. Mine is more mechanical, which is

even if you had infinite information sources, it would not be optimal to consult them all predicate to acting. I see you as making a slightly different point, which is you might increase uncertainty, the more you dig into something, by understanding how complex it is.

Primavera De Filippi

- Yes, yes. She seems to claim that there can be no knowledge without a particular degree of trust in something. And, I actually think I agree. Uh, I don't know if I agree in the sense in which she speaks about it. I think I agree just because, I see that even if I have confidence in things, that confidence emerged from trust, and eventually there is trust all the way down if you like, trust or confidence all the way down, but in the end, there has to be a level of trust. So yes, there is a point at which I cannot, I cannot accept any knowledge if I don't trust at least myself. But I think the way in which she says it is very different from the way in which I perceive it, in the sense that I think if we stop at the first layer, to me, you cannot have knowledge without trust because that knowledge comes from confidence. But then if you look at that confidence, where does that confidence come from? It comes from either trust or confidence. And so you can go all the way down.

Ori Freiman

- It is the problem of the skeptic and those that emphasize, you know, traditionally they said, okay, I'm skeptical of what you say, let me see the results myself. So yeah we copied their discussion, but in new terms, it's nice.

Michael Heidt

- Yeah. I, I think it might make sense to return to systems theory once again, because just my observation always is that scientists are really careful to point out that what they are producing cannot inform actions, or isn't intended to inform actions in any way. Ask a sociologist, what is to be done about society, they've often quite forcefully said, I don't know, I can't say, that's not my job, uh, like in informed politics or something like that. And it's not a completely different level. Um, yeah. Maybe systems theory can help here because what science then produces is just different distinctions, different ways to describe the world, different points to be made, which then reduce complexity which you have to process through actions. This also points to this problematic in economics. Because often I feel there's an even deeper level of uncertainty, so I can see to us dealing with uncertainty as well. Um, but often for me, this uncertainty is on an even more basic level. Like if I trust my intimate partners, how destroyed would I be, if they misuse that trust and cheat on me, like I can't quantify it. And it might destroy me as a person and destroy the basis on which I would be able to deal with this complexity at all. And in economics, not all of economics, but you often have this universal medium like money. If

I, uh, If I'm confident the markets will grow up or I will go sideways. I could sell a lot of options right now. And well, if the markets are going to form well, that's the risk Iuy take. And then I will owe my broker $2 million or something that I might throw up. But, um, even this catastrophic event is neatly describable. And, um, I think a lot of these other systems lack this universal medium. So that's kind of a structural difference for me with this, uh, economics example. It's like having a deeper level of uncertainty having to decide which distinctions to apply and how to frame these actions.

Eric Alston

- I think, I mean, there's a commensurability problem that most economic theories suffer too, once they become mathematical and certainly it's of necessity, a limitation of empirical data that there's counting involved. And so the scope of that type of analysis to relatively narrow questions, completely agree. And this is also why I think I failed my PhD comps in econ grad school because the theory was not my cup of tea. But I do want to plant the humble flag that there is a subset of economists that view economics as the science of voluntary human exchange and all of that might entail, including the rules governing that exchange. There's at least a subset of economist that I, I, I like to believe I have a slightly richer understanding of human nature than zeros and ones and maximizing a utility function, subject to a preference set and budget constraint, but point well taken, I would say most of the field of economics is very subject to the critique you just made. But I think there are nonetheless interesting thinkers who come from the field of economics who are grappling with questions of uncertainty who are grappling with questions of trust and not all of them are as enamored with math as the editors of the top economics journals are these days. That's for sure. Um, so point point well taken.

Primavera De Filippi

- I would say that most game theory is actually dealing with trust. It’s all about like, do I coordinate or defect, do I trust that the other will coordinate or defect, and so forth.

Eric Alston

- Yeah. Although so that being said, the game theorist needs a payoff function that's either numeric or binary. Otherwise their, their problem becomes completely intractable to resolution, like a game theory paper, where the payoff function is in complex numbers, let alone isn't binary or something that can solve the results to, then you reach a big problem. And so, like I love game theory. I think the prisoner's dilemma is illustrative of many, many, many situations in real world context. So I don't mean to belittle that. But I do nonetheless think game theory can still be subject to the commensurability and quantifiability critiques that I saw as suffusing part of Michael's statement. That's not all you were saying.

Ori Freiman

- If we go back to, if you go back to system theory, there are those rational choice accounts of trust. We have those kinds of models, for example, Bayesian models. I don't know if I will buy it, but those accounts obviously exist and they are useful.

Eric Alston

- Economics often gets pilloried because again, an assumption to make things tractable to analytical resolution is typically the rational choice model, but most economists don't believe people behave rationally all the time, but is there a better, if you were to say, I have to throw a dart, is there a better prediction for how people will behave all the time? Or to put it differently, at least 51% of the time. Do you think people rationally maximize their expected benefits when subject to a decision? And there are plenty of people who think, not even that, not even at least 51% of the time, but I think a humble defense of the rational choice assumption is if nothing else out there is defensible as against that standard as a single prediction, that's why one discipline uses it, uses it like a blunt tool most of the time, but,

Ori Freiman

- Okay I completely agree. But if we focus on the agency not of the individual as a rational choice, because obviously we humans, nobody would argue we are rational. But if we try to make some kind of a decision taking for one organization for an algorithm and we want to program it. So we have no choice, but to make it a rational choice in terms of if you want to code trust. So it would be like a rational choice to trust. But in the sense not of an individual human, but in the sense of, uh, some other kind of agency and that's thought provoking.

Primavera De Filippi

- So actually we were having this discussion last week or something, uh, about the social credit system and it was kind of related. What are the ways in which you can evaluate, and there are many ways in which you can rank whatever, like assess things. Um, the usual way that we've been using in traditional times, uh, is actually expert evaluation, right? Like, so you have people that you consider to be more knowledgeable about a particular topic and they are gonna evaluate a particular thing. And then all of a sudden we have the internet and crowdsourcing. And then the question is like, well, now we can actually have peer to peer or crowdsourced evaluation. So the evaluation, we don't need experts anymore because we have the crowd and the agglomeration of the crowd might actually give a better result, mauve less biased, who knows. The interesting thing is that once you start having multiple evaluators, then you need at least a metric. You need at least a standardized metric system. Because if I'm ranking people and I have a different metric than you, then you cannot aggregate, then it doesn't make sense. And then finally

the third one, which is like AI and computer based assessments, which is actually inherently numerical. So it's not even just like, you need a common metric. It could be ABC, like could be good, bad, medium, but now you need numbers because the algorithm only understands numbers. And so all of a sudden, as we want to rely more and more on technology or collective aggregation, we're actually moving away from qualitative assessment. Because it's not that we don't like them. It's that the new tools that we're using cannot afford it. And so there's this huge, huge tendency to datafication and metrication, et cetera, just because we're changing the tools by which we are trying to make assessments.

Michael Heidt

- I think that's actually an interesting point in the paper because it just gets at, how you have to highlight and foreground this continual process of formalizing and representing these distinctions through algorithms, so not try to have the correct way of a formalization, but to establish communal awareness of how to continually negotiate these formalizations.

Eric Alston

- I'm reminded of early arguments I made against using blockchain through land titling problems in the developing world. And to me, my argument always was that it 's an initial state problem. So until you resolve some very complex adjudication and possibly force a lot of conflict by resolving with finality all lurking property disputes, it's very hard to have a good, like sort of trifecta identity mapping between a piece of property, an individual, and the set of social relations that that mapping actually entails. I'm more sanguine about pilot projects in Singapore, in Sweden because the iterative development of our land registries is one that's approaching the level of commensurability that Primavera noted was a predicate to sort of algorithmic comparison between things that you've given labels, let alone numerical definitions to, and so this will always be an ongoing problem with respect to linking algorithmic processes to real-world outcomes. I just think it's unavoidable, but I also don't think it can't be overcome. I'm not in that camp, but it's a challenge.

June 17 — Nguyen (forthcoming): "Trust as an unquestioning attitude"

Attendants: Wessel Reijers, Ori Freiman, Judith Donath, Quinn Dupont, Victoria Lemieux, C. Thi Nguyen, Primavera De Filippi, Balazs Bodo, Eric Alston, Christopher Wray, Charles Nesson, Sankalp Bhatnagar, Michael Heidt, Juan Ortiz, Simona Ramos, Philemon Poux

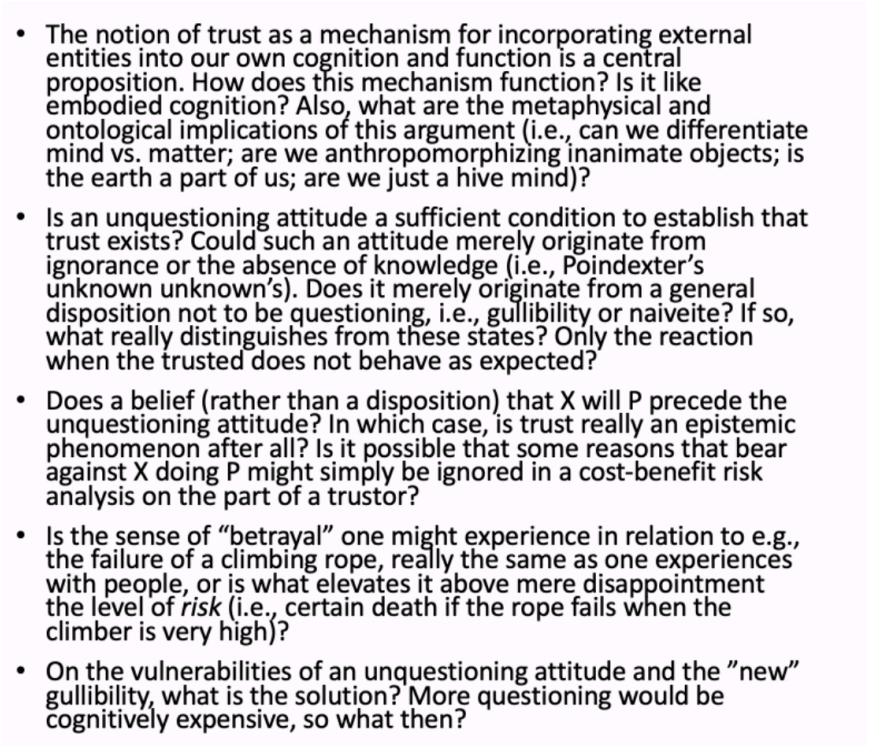
Discussion Notes:

Thi

- I always teach Baier's trust and antitrust. It was about vulnerability and it seemed like the language of vulnerability was such a natural language for me, when I was groping to talk about our relationships with things like Fitbit and Twitter's metrification system, but obviously Baier and what follows from her that language is so specifically about human agents. Once I plugged in my laptop to project to a screen so it changed the resolution. And when I unplugged it, all my folders were in the wrong place and I felt such incredible rage. And then that's when I started thinking about how you could be betrayed by objects.

Victoria

- You're basically building a pathway to thinking about trust in non agential technical artifacts, which is a challenging aspect of applying the concept of trust to artifacts as opposed to trustworthiness. I find a lot of what you say in your paper quite compelling and novel. You start off by saying that most existing accounts of trust, and you just mentioned Baier, are based on the assumption that trust is agent oriented. Agents that are sentient and think about the interests of the trusting party and behave in accordance with those interests or not, um, as just one example and that, uh, so the trustor must describe some complete essential state to the, to the trusted. Trust is only directed towards things that can bear agential states, people or groups of people, including nations and corporations. And you point out that in some cases it might be possible to stretch the concept to incorporate complex technical artifacts, but it really doesn't encompass dumb artifacts or natural things. You think that there's another way to think about trust that does allow for that.
- An unquestioning attitude is not the only form of trust. It exists among other forms. Um, but this unquestioning trust is an unquestioning attitude in which the trustor steps back from a deliberative stance. They are no longer monitoring, challenging, checking in or questioning the trusted. You say that trust entails a process of integrating the trusted, so people and resources into our own cognition and action or functioning. There is cognitive efficiency but also vulnerability. The unquestioning attitude plays a role in settling the mind and reducing the demand for cognitive resources.
- Trust is about betrayal, and what's happening psychologically is a form of alienation from ourselves because we bring the trusted into our cognition and our action. It's not mere reliance because we might still need to rely on something that's not particularly trusted. Our memories, for example, because that's all we've got. All forms of trust we have are about expanding our own agencies, so integrating the external world. So they all share that as a kind of a common foundation.
- Critical reflections:

- 

Thi

- To your last question I don't think there is a solution. I think that our basic dilemma as cognitively limited beings with finite processing power and finite time and finite resources facing the world that is vastly too large for us. One, the only attitude that can get us through is trust and two, every trust makes us vulnerable. And three, we don't have enough cognitive resources to secure our trust. And that's our basic epistemic dilemma as finite beings. I also don't think that total intellectual autonomy as the ideal can be the proper stance in the modern era. So my life in epistemology was changed by reading a book, the great endarkenment, which basically argues that the essential epistemic situation of humanity post-enlightenment is one in which not only is it that science is vastly larger than a single human brain, but no practical argument that is science-involved is understandable by any single agent. And furthermore, the harder criteria to swallow that both Elijah and I believe is that not only do you not have the resources to do all the intellectual processing yourself, you don't even have the resources to properly identify the right experts.
- The question we have now is what the hell do we do? How do we survive as cognitive agents who need to be vulnerable, who need to exist in this realm, in which information is far too large for us. And yet at the same time, we can't just trust everything. So the question of how we manage our vulnerability and how we secure our trust, knowing that it's not possible to secure it perfectly, that I think is the question for us.
- I have a new paper called Transparency is Surveillance. Transparency asks experts to explain themselves to non-experts. Um, but to do that, they have to change their reasons. And, uh, this leads them to either make stuff up or in my formulation actually changed

the way they were using reasons someone could justify themselves better to non-experts for example, legislators or oversight. So I ended up thinking something like there's this fantasy that people have, you have pure transparency and pure transparency involves the public in general court in general aggregate non-experts having oversight over experts. And that's the condition that you have when you try to approach things by eliminating all trust. Uh, but that leads to the squashing of expertise, but of course you can't get rid of transparency because then corruption bias. So with trust and transparency, there's a slider and you often have to choose where you are along that slider. And the more transparency you have, the less corruption bias you have and the less you trust experts to be experts. And the more trust you have, the more you get to unleash the power of experts. And the more you let the possibility of total shit into your system. And there's ways we can make the compromise a little better, but I think the essential situation is one of compromise, and if you try to solve perfectly for understanding and you try to eliminate trust from the system, then you lose most goods of social coordination. Complete transparency leads to the significant undermining of expertise.

Christopher

- Have you considered relating this to System One, System Two and Thinking, Fast and Slow?
- And is there a hierarchy of information processing or heuristics we can use as cognitively limited beings?

Thi

- Trust enables enormous interpersonal cooperation, but also permits extreme abuse.
- I'm suspicious of System One, System Two stuff, and I found my suspicions really well illuminated in the recent work by Mercier and Sperber.

Eric

- Just yesterday, a tree trunk that I consistently use to cross a very raging mountain river to access a portion that I scramble regularly was taken out by the whitewater. Had that happened while I was on it, I would be dead. And I was on it three days prior to crossing this really gnarly river. And so in a sense, I was definitely in every instance and in order to safely cross it, I was putting to the back of my head, any such concerns about that inanimate, non man constructed object being taken out.

Thi

- The standard, Cindy Goldberg view is that we only trust an object when you're trusting the people behind them. I think this can't be right. Sometimes the object departs from the manufacturer's circumstances. So when I buy a new rope, I trust the manufacturer. When I decide to use my two year old rope after some assessment, my trust is more complicated.

And I find this important because I think a lot of the things that people are trusting in emergent social media network architectures were not designed in right. They're emergent features.

Quinn

- So unquestioning attitude, you got me on the unquestioning, the attitude part though is where it gives me a certain amount of discomfort. And when I see this descriptor I'm initially drawn to habit and, you know, specifically John Dewey's articulation of habit, and my question is why not habit as opposed to your attitudinal approach?

Thi

- So habit for me is a subcategory of attitude. I think the desire to push it into a system one system two or a habit framing still smells to me like a very individualistic approach to cognitive processing. Say you buy something and it's clear how it works. And then in the middle of the night, it updates and you've consented to version one, but you have not consented to version 1.5, right? So the target of my trust in a case where technology auto updates and changes the way it works, I'm not sure how to flesh that out with the notion of habit. I think that way with Google search. You're outsourcing a part of your attention and when Google search changes its algorithms, that's a change in the context of your trust that I can't really talk about as a, in terms of habit explaining the central content. And another reason I don't want to use habit is because I think in many cases, um, putting it as habit still imagines that you have some sense that you're trusting something. I think my mother has no idea what she's trusting when she opens up a computer. Like we've told her to do certain things, but the relationship is different. It's not a sufficiently internalized account with habit for me.

Ori

- Can you see some kind of a role for the social norms that govern the process that ends up placing trust in technology?

Thi

- Trust is when you suspend deliberation. We do it for all kinds of things; a lot of times we do it without realizing it. Another question, what's the criteria for suspended deliberation? Really complicated, given that you don't have the expertise in the things that you're trusting.
- There's this incredibly rich literature about dataism and the movement towards metricization that's happening in sociology. All of this is about how the transition from qualitative information to mostly quantitative information moves the kind of information we pay attention towards, towards the things that are standardizable and observable. And all the information that gets lost because of that. All are part of this really complicated

story about what happens when trust meets large scale institutions and we start getting the metricization of the reporting of success.
- Scott, Seeing Like a State. Ted Porter, Trust in Numbers. Wendy Espeland, everything, but especially Engines of Anxiety. Susan Star, everything, but especially Bowker and Star, Sorting Things Out.

Wessel
- In terms of Heidegger on tool use, is there really a lack of deliberation where we are in this mode of dealing with the tool unquestionably before it breaks down?

Thi
- This paper was trying to think of trust in specific things that are identifiable entities. And then I think you're talking about something really important, which is trust in whole systems. My view was never that you have this fixed attitude where you never deliberate about certain things. It's that the attitude that we take up of trust is non deliberation. And then sometimes we deliberate about it. Uh, and, and then trust in this sense is suspended for the moment. And we do this all the time.

Charles
- I'm thinking of trials and juries and the role of trust in the process of deciding on, delivering and accepting a verdict.

Thi
- I'm very interested in the way large-scale institutions create trustworthy-seeming systems. There's a great book, Seductions of Quantification, about the UN's indicators about, you know, which countries are doing well in sex trafficking in which countries are doing well from a human rights development. It's this process where you take all these fuzzy, complex inputs that no one's sure about. And then you do this mechanical process out of public sight and it's closed doors and then you generate this ranking system. And then people instantly trust the thing that's generated and then act on it. It's an attempt at objectivity, even though it's unclear how objectivity got in the system.

Primavera
- I love trust as this second-order act of not questioning. But there seems to be a missing reference and analysis between how this unquestioning attitude relates to confidence, which in the Luhmann sense is about taking things for granted that they will work in a predictable manner, which is also this notion of unquestioning attitude. So does confidence fit within this framework, or is it something else? And also what about familiarity, can't that also lead to unquestioning attitude? There might be many reasons to reach this unquestioning attitude, and not all of them might be related to trust.

Familiarity might actually lead to unquestionably doing things because you are so familiar, not because you're confident or trusting, but familiarity is leading me to this unquestioning attitude. Another is knowledge, expertise, and experience – bringing confidence to an extreme. I'm unquestioning that the sun will rise tomorrow, I will act on that, I'm not trusting but just confident. So bringing familiarity or confidence to the extreme reaches this unquestioning. And trust in the extreme acts the same way. They all lead to it. But then regardless of which path you took to get to unquestioning, is it that you will always feel this sense of betrayal, whenever unquestioning has been broken? Or only when the source of unquestioning is trust?

- My understanding of the difference between confidence and trust is that when expectations are broken, in the case of confidence you are blaming others or the world. So if the ground falls under me, and I was confident in the ground, I blame the ground. Whereas when I trust and my expectations are broken, I will blame myself because I willingly put myself into a position of trust. Your paper showed that when you internalize something external as being part of you, then if that fails you will blame yourself. And you fall back into the same attitude as a trusting attitude, where something went wrong and you blame yourself because you consider it to be you. If the phone is an extension of yourself and it fails, you blame yourself, instead of blaming the phone producer. So maybe this justifies why sometimes we feel betrayed about something we had an unquestioning attitude towards, which is a different reasoning than the paper's. The difference is in how we relate to broken expectations.

Thi

- I think all the pathways are about this sense of betrayal, but I haven't thought of confidence enough. Please send me readings on that!

July 1 — Jackson & Sunshine (2007): "Public Confidence in Policing"

Attendants: Ori Freiman, Jon Jackson, Beatriz Botero, Morshed Mannan, Eric Alston, Wessel Reijers, Paula Berman, Primavera De Filippi

Discussion Notes:

Jon Jackson: My PhD looked at fear of crime, and how people make sense of crime through low level signals of disorder, social cohesion, collective efficacy, kids hanging in the streets, a sense of shared values, etc. And this paper extends that to how people think about the cops. The argument is they don't think of police in terms of risk of crime. But as symbols of social order and their perceptions and concerns around social cohesion. Even taking into account perceptions of the

procedural justice or effectiveness of the police. Is the neighborhood orderly, cohesive, trusting, acting for the common good.

Beatriz Botero: Exactly, and the conclusions are that people lose confidence in the police when they see shared values, not safety, deteriorating; and that if police seem to represent community values, the public has more confidence in them, almost independent from actual crime, and they have more legitimacy. Public confidence in the police is also related to authority, and not necessarily their force. This is driven by how much they're trusted, which is related to how fairly they are perceived in their procedures and in their outcomes. It's about how they convey values and the manner in which they exercise their authority. My questions are: how has your thinking and research evolved? And how does more diversity change the public perception of police given its effect on social cohesion and shared values?

Jon Jackson: First, legitimacy is about a relationship, a dialogue, between a power holder (institution) and the people it's making moral claims of power over. And there are expectations that define what the power holders should do to be seen by the audience as moral, just and appropriate. It's about people's perception and acceptance. Second, as a psychologist, thinking about legitimacy without some aspect of consent makes no sense. Police represent policing, and so in an area which seems to police itself, thus having collective efficacy, and which doesn't have disorder, and thus the signs of the failure of police, the police get some credit for the informal policing working, as the formal institution of policing. They seem legitimate.

Wessel Reijers: I like that you're trying to empirically capture the concepts we've been discussing. Given that there might be distinct understandings of legitimacy (de facto, normative), where there might be reasons to consent to authority vs good reasons, how can you translate this into some kind of empirical research? Usually empirics can gauge people's preferences but not their reflection on those preferences.

Jon Jackson: I don't see the relevance of legitimacy to blockchain. If legitimacy is de facto: blockchains exist, some people use them and some people don't, they've become a social fact. To me legitimacy is about the psychology of power differentials. And the dyadic relationship needs to be a normatively grounded obligation to consent and not just compulsion for it to be legitimate. In political philosophy, there's this distinction between the empirical concept of legitimacy and normative. Example: a group of experts could design the perfect criminal justice system. But a community with different values might legitimate an institution that people/experts outside that group would not. It's about local expectations.
I'm trying to understand how legitimacy applies to some aspect in society that doesn't actively make demands on you?

Primavera De Filippi: Legitimacy is interesting in blockchain because everything is voluntary. And we're wondering if increasing confidence can increase, rather than eliminate, trust. And can trust also be a source of legitimacy: if I trust the person that has power, I might consider their actions and decisions more legitimate. We've conceptualized trust as voluntarily delegating power to someone because I think this person will act in line with my own interests and values. While I voluntarily give trust to someone, legitimacy just is, as in there is power that is being held by an external system. And if I didn't decide to put myself in this situation of vulnerability toward this actor that has power over me, and I can't get out, do I consider this legitimate? Trust is voluntary, and legitimacy is more about consent. So to what extent does the level of trust or confidence inform the perceived degree of legitimacy.

Morshed Mannan: Even accepted as a social fact and whether you use it or not, there are stakeholders in the operation of a blockchain, and there is a question of whether they are able to exercise their will on others using the blockchain. Unlike in a normal policing environment, where there is no option to exit, here you can exit. We wonder how legitimacy can contribute to improving this process of consent and approval you describe, as well as thereby making governance better. So do you think legitimacy only becomes relevant when the institution is able to prevent exit, or whether legitimacy is also relevant when you can voluntarily join a system? So maybe like a public system versus a club.

Jon Jackson: People can exit from policing. Most interactions with the police don't involve the threat of naked force, and proactive voluntary cooperation is a key part of what we study. Legitimacy has to be about hierarchical power relations, when institutions are making a demand on you, implicit or explicit. You can comply instrumentally or normatively. So if someone complies with the law or cooperates with the police, for it to be truly called legitimacy, it has to be independent of any other motivation to do the act. Legitimacy is willingly following rules, independent of the content of those rules.

July 15 — Tonkiss & Passey (1999): "Trust, Confidence and Voluntary Organizations: Between Values and Institutions"

Discussion Notes:

Wessel: The relationship of voluntary organizations to the issue of trust is highly problematic
Trust and confidence

- Trust: ethical relations not conditioned by external framework of controls
- Confidence: relations secured by contract or other regulatory forms
- Potential conflict: confidence-based measures formalize trust

Trust in voluntary organizations

- Seligman: voluntarism crucial to understanding trust (in sense of shared values, separate from relations of confidence)
- Trust becomes problematic when transposed from personal to institutional level
- Generalization of trust happens through secondary organizations, between family and state
- Fukuyama: network of moral communities, critical to organizational efficiency
- Yet: not only simple model, but increasingly complex social and economic institutions
- Voluntary organizations have more of a stake in governance, and their operating environment has been more regulated
- Tension between 'doing good' and 'doing well'

Why trust matters

- Linked with values of honesty and fairness
- Charities are highly trusted, but also face a crisis of trust
- Factors: (i) trust basis for voluntary organization; (ii) trust represents public goodwill (time and money); (iii) trust hands a political license to operate (legitimacy?)
- Empirical research: focus groups and survey

Findings

General public

- Trust closely linked to values (caring functions); helping those in need
- Ambivalence towards the way charities work as institutions
- What is the role of charities vis-a-vis the state?

Government and institutional funders

- Funding relations increasingly conform to standard contracts
- Relation between public funder and association one of transaction
- Targets, output, audit, regulation (formalization)

Business

- Independence from business
- Corporate philanthropy as symbolic credit
- Move towards business oriented two-way street of social responsibility (civic values => market values)

Users or beneficiaries

- Conflicts about influence of users and professionals, leading to formalized systems of decision making
- Language of rights to raise questions of internal democracy; hardening of trust relations

Conclusions

- Trust relations based on voluntarism – voluntary association
- Trust is linked to shared values
- Trust relations are distinct from confidence relations
- This in particular generates difficulties of trust in the voluntary sector

- A different understanding of system trust; usually there is no reflection on the type of system involved; e.g., spheres of family, state, and secondary organizations

Questions

What is the link between voluntarism and confidence/trust in blockchain communities? Which actions are strictly speaking voluntary? Blockchains are inherently formal/procedural, does this mean that they by default conflict with trust? Or crowd out trust relations?

Fran: It depends on what you are relying on, in any kind of organizational relationship and interpersonal relationship. What is the basis of the claims you make within that relationship and what is the basis of your expectations of others? And I think that's where trust and confidence can co-exist. But the more contractual, the more regulated, the more conditioned an interaction is, the less space I think there is for trust... If you think about childcare, the difference between having a member of your immediate family, perhaps the grandparents of the child, look after your child for free. Usually this is based on a relationship of trust, fundamental and unconditional, and at a very low cost, you don't have to pay for it. And forms of childcare provided through those intermediate associations or through state forms, increasingly the relationship has to be one of confidence, but you would never want to have your child cared for by someone who at some level you didn't trust. But trust is more an effect of the initial confidence relations. Most interactions sit somewhere along that continuum between absolute trust and watertight, contractual confidence.

Primavera: I agree with the paper that trust is reinforced by and also generated within an organization, whereas more institutional arrangements rely on confidence and generate more confidence... But shifting from trust to confidence, does it always weaken trust? Can providing confidence, through guaranteed execution and transparency, enable interactions that would not otherwise be possible, which creates new opportunities to build trust. A proper institution has elements of trust and confidence, which can actually reinforce each other. But bureaucracy might just have confidence. Vicious circles of institutionalization can eat up trust.

Fran: Yes, consider workplaces: strong relations of trust built up on social interactions, as well as bureaucratic forms.

Vashti: How does one reconcile the movement from personal to institutional trust with the need to mitigate inherent risks?

Fran: What a non-expert might see as extremely risky could be mitigated not simply through technological fixes or forms of knowledge, but through some kind of social stake and the stakeholder model.

Morshed: How do we create solidarity, which usually depends on repeated interactions, in a pseudonymous environment among strangers... A system geared toward solidarity would be geared toward voice rather than exit. A lot of DAOs and blockchain actions are geared toward exits. Part of the reason solidarity is created in legacy stems is because there are fewer options for exit.

Primavera: We usually talk about legitimacy when there is an authority or coercive power, and the question is assessing the legitimacy of that power, do we accept it or just submit to it. And in blockchain you have full freedom of exiting cheaply. And because of this, legitimacy becomes fundamental, because I can leave to another system if I don't see it as legitimate. Here there is no single centralized authority, it's a distributed system. It's not authoritarian, it's not coercive or obliging anyone to floor a decision. But legitimacy is key because of the easy exit. So legitimacy also makes sense perhaps in a situation where there is no specific authority or coercion, and it becomes a necessary force for a system's sustainability, because as soon as it loses legitimacy people will leave.

Fran: Why would people see blockchain as illegitimate?

Primavera: A flaw in the code, some having opportunities to manipulate the system, some disputed upgrade to the code (or maybe I don't like the upgrade but I still see it as legitimate so I accept it). Or maybe the government starts saying it's illegal, so it's illegitimate from the outside at least… Do you see blockchain as a confidence machine? Harming opportunities for trust relationships?

Fran: Given it lacks so many of the forms of protection and accountability that are conventional in organizational life and certainly in economic life, confidence seems to be very important. But then I'm mindful of all the social noise that goes on around the operation of the blockchain, I mean, all the forums and all the chats and all the social media and, and so on, which seems to be about generating something else. Clearly you need a diverse language to get at the different elements of this.

Primavera: Blockchain is designed at least to build confidence in the system, but that cannot exist without trust underneath, that enables confidence to emerge. And that confidence could perhaps then help to establish more trust relationships. It's this fractal or concentric system, where confidence in a particular system can contribute to enhancing trust and confidence in another system. And your paper shows we cannot just generalize and talk about systems, uh, without actually looking at the typology of the system.