Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity* Article 17

February 2020

A Metacognitive Approach to Trust and a Case Study: Artificial Agency

Ioan Muntean
UNC Asheville

A metacognitive approach to trust and a case study: artificial agency

Ioan Muntean
University of North Carolina, Asheville, NC 28803, USA
Email:imuntean@unca.edu
http://imuntean.net

2019-09-05

Abstract

Trust is defined as a belief of a human H ('the trustor') about the ability of an agent A (the 'trustee') to perform future action(s). We adopt here dispositionalism and internalism about trust: H trusts A iff A has some internal dispositions as competences. The dispositional competences of A are high-level metacognitive requirements, in the line of a naturalized virtue epistemology. (Sosa, Carter) We advance a Bayesian model of two (i) confidence in the decision and (ii) model uncertainty. To trust A, H demands A to be self-assertive about confidence and able to self-correct its own models. In the Bayesian approach trust can be applied not only to humans, but to artificial agents (e.g. Machine Learning algorithms). We explain the advantage the metacognitive trust when compared to mainstream approaches and how it relates to virtue epistemology. The metacognitive ethics of trust is swiftly discussed.

Synopsis

We argue for the importance of *metacognitive* requirements on trust. Trust is a two- or three-place predicate and it implies that H, *i.e.* the human trustor, expects that an agent A (the 'trustee') 'will take care of the things'. (Baier 1986; Jones 1996; Carter 2019) H has a belief or a positive disposition to think that A is going to perform a set of actions in the future, on which H depends to some extent.

When is H's trust in A rational? How much does H need to know about A in order to entrust it? As trust always incurs uncertainty and a non-negligible level of risk and uncertainty, a more formal analysis based on Bayesian epistemology is more adequate.

A deflationary view about trust is to claim that trust is just a form of reliance. We trust (or distrust) friends, relatives, experts, some communities, some

institutions, scientific communities, and some domesticated animals, but we rely only on some artifacts or natural entities. Trust and reliance have a common ground: we trust humans to behave predictably, in a sense that regular behavior applies to other agents. (Hollis 1998) The case of predictability of the behavior of artifacts is enticing philosophically. Artifacts, including artificial agents, sit between natural agents and human agents. One way to differentiate trust and reliance is to admit that agent A needs to have some epistemic and moral competences. We take this difference further: we advance a metacognitive argument for trust: A needs to display a multi-level type of knowledge, which suggests a virtue epistemology approach (Sosa, Greco, Carter) The present paper adopts a naturalized virtue epistemology by integrating traditional epistemology into the theoretical framework proposed recently in cognitive science (Fleming, Daw, Proust, Meyniel, Timmermans, etc.).

We argue that rationality of trust entails a more throughout analysis of H's and A's competences and that a metacognitive approach to trust adds significant mileage to the existing arguments on the rationality of trust. Trust of H in A (human or artificial) implies a process of rational deliberation about the competencies of A. The cognitive agent A, be it artificial or not, has a set of goals, makes some choices, and is able to calculate costs and gains of its own action. A set of metacognitive requirements can be added to the cognitive agent A: confidence level, ability to detect its own errors in modeling the world, or more elevated metacognitive competencies such as: self-reflection, humility of judgments, ability to suspend judgment, or consciousness.

Further, we claim that the requirement of A being rational can be couched in terms of probabilities, conditionalizations, and computational processes. Let us assume that a rational agent A is more trustful than a less rational agent A'. In the metacognitive framework, we think that A' can be characterized as having fewer (or none) metacognitive competencies in comparison with A. In the line of the naturalized approach preferred here, we can relate rationality to some metacognitive processes that A is able to instantiate. The best framework to synthetize these ideas is the Bayesian epistemology. We start with a simple model in which the world has two states only and A perceives the it with a given amount of noise and distortion and makes a decision about the state of the world. After discussing briefly the concepts of 'accuracy' and 'confidence', we propose a cognitive model of trust based exclusively on accuracy and a metacognitive, two-level, model based on confidence. Confidence is taken here as an independent computational process happening in human agents (Fleming&Daw, Timmermans, Meyniel) or in some artificial agents. In the confidence definition of trust, we assume that the probability of accuracy of A's judgment increases conditionalized upon the calculation of A's confidence (when confidence of A is offered to H as an independent parameter of the decision-making process, similar to a statistical result). Some authors think that confidence can be a simple statistic summary of the judgment or can be a property of the populations of neurons used in the judgement (Timmermans). We conclude the paper by discussing the case of artificial agents based on machine learning in which different procedures to avoid overfitting can instantiate some of the metacognitive re-

quirements: suppressing of data noise, and model uncertainty (Gal, Mackay). A brief digression in ethics about metacognitive requirements on trust concludes our paper.

1 Introduction: Trust matters

When is it rational, or irrational, to cooperate, or to not cooperate, with other agents? How much do we need to know about other agents in order to be ready to cooperate with them? Is it morally right (or morally wrong) to cooperate with this particular agent? Moreover: how will cooperation look like in the future societies?

Most answers to these questions involve in a form or other concepts such as 'reliance' and 'trust,' which have played historically a constitutive social and political role in human societies. (Gambetta, 1988; Cook, 2003; Hosking, 2014) Nevertheless, in the current political and social climate, we are more and more concerned about the *erosion* of trust in institutions, experts, democracy, religion, education, etc. We are currently witnessing the downfall of trust in science and in technology. A future society with isolated individuals losing trust in their peers, reclused to computer games, virtual reality or social media is equally the future of a society with little trust or no trust.

Is the social cooperation in the communities of the future going to be based on trust, as it used to be during the progress of humankind? Or, on the contrary, will we witness a trustless society dominated by individuals unable to build trust and to rely on their fellow citizens?

Trust in an informal way implies that H (a human trustor) hopes or expects that A (an agent, the trustee) "will take care of the things" as H entrusted A. (Baier, 1986; Jones, 1996; Carter, 2019) This is called by some the doxastic view of trust, in which H has a belief that A is going to perform a set of actions in the future. As trust always incurs uncertainty and a non-negligible level of risk, it can explain why cooperation is always risky.

Like trust, reliance is a cognitive attitude towards artifacts, bodies of knowledge (scientific theories, religious doctrines, traditions, ideologies, etc.), institutions, communities, etc.[1]

The mainstream philosophical literature views trust as 'more' than reliance, and some requirements can be imposed on the trustee A: personhood, trust-

[1]Some people claim that we rely (or not) on bridges, cars, buildings, technologies, scientific theories, institutions, legislations, religions, etc. but we do not 'trust' or 'cooperate' with it. We trust (or distrust) friends, relatives, experts, some communities, some institutions, scientific communities, and some domesticated animals. We cooperate with these agents and not merely rely on them. Trust and reliance have a common ground: we trust humans to behave predictably, in a sense that regular behavior applies to other agents: some animals, some phenomena governed by natural laws, etc. (Hollis, 1998) The case of predictability of artifacts is enticing philosophically. Artifacts, including artificial agents, sit between natural agents and human agents. Does a team of surgeons 'cooperate,' 'trust,' or 'cooperate' with a robot surgeon? Do doctors trust the robot surgeon or merely rely on it? What do we mean when we say that the robot surgeon 'complements,' rather than 'replace,' the human surgeons? Similar questions can be raised about unmanned vehicles, lethal weapons or diagnostic systems.

worthiness, moral responsibility, goodwill etc. This leaves us with a restrictive concept of trust: in this view A is endowed with human agency, hence the *interpersonal* nature of trust.

1.a Main aim of this paper

We do not endorse the idea that trust is only inter-personal—we prefer a more general concept of trust to include non-individual or non-human agents A. This paper focuses on the epistemology of trust: its rationality, its metacognitive aspects, and the way it can be generalized to other types of agency, especially artificial agency. Our main aim is to argue for some common ground between the way we could entrust artificial agents and the way we entrust human agents. The best theoretical framework to bridge trust in artificial agency with existing approaches to interpersonal trust is a naturalized version of virtue epistemology that employs probabilism, conditionalization, and tools from Bayesian epistemology.

We are interested in requirements imposed on A such that we build the 'right' trust in reliable technologies of the future and a precautionary stance towards those technologies which are dangerous. How do we build trust in AI (artificial intelligence) technologies and other algorithms (especially Machine Learning) as 'autonomous artificial agents' (hereby, AAA). We argue that trust needs a metacognitive component and that Bayesian epistemology can be successfully used in this respect.

1.b Philosophy and trust

There is a rich literature on trust in social science, psychology, and philosophy. Philosophers look for a genuine concept of trust to contrast it with folk concepts of trust, so much of the philosophical analyses are based on some restrictions on trust, with a normative load. (Nickel, 2017)

Most restrictive definitions of trust follow a template like this: "To trust A means to rely on A and to believe that A has X". The requirement X is some superior cognitive or moral capacity. It can be interpreted as a disposition, as a competence, or a high-level skill. It is a good idea to interpret trust as a set of constraints imposed on a weaker concept such as reliance or some naïve forms of trust. We suggested that philosophers prefer to talk about inter-personal trust, when A is not 'something', but 'somebody.' A must be a person with her agency, desire, goodwill, moral compass, moral responsibility, high level knowledge, social profile, etc. Jones (Jones, 1996) 14 writes: "One can only trust things that have wills, [...] although having a will is to be given a generous interpretation so as to include, for example, firms and government bodies."[2]

[2]Hawley explains that one can trust somebody to look for a vase, but one only relies on a shelf to hold the vase. (Hawley, 2014) If I break the vase, its owner, who entrusted me, can be disappointed in me, betrayed by me, or demand an apology from me, although the owner will not feel the same about a shelf that was supposed to support the vase, although the owner may have some similar feeling towards the designer, or the manufacturer, or the assembler of

2 Virtues, knowledge, metacognition, and trust

One way to advance a metacognitive argument for trust is to define it in a multi-level epistemology, and then naturalize it. Our choice is here for virtue epistemology, which is considered by some a non-standard, non-traditional, epistemology. We then argue for a naturalized version of it in the metacognitive framework.

2.a Levels of knowledge and epistemic virtues

In virtue epistemology, knowledge is a form of competence or achievement. (Sosa, 2007, 2018) A competence is the disposition to perform well in a given domain. E. Sosa postulates two levels (or 'grades') of knowledge: at the lower level (called 'animal knowledge') the knower does not have an epistemic perspective on her belief, but she exercises a reliable competence to believe. In Sosa's view this level requires *apt belief*, a belief that is "correct attributably to a competence exercised in appropriate conditions" (Sosa, 2007, 2018) 93 In this view, competence accounts for why knowledge is superior to a mere true belief. Aptness does not require a reflection on one's own.

To have high-level knowledge ('reflective knowledge'), the knower needs a perspective from which she endorses the source of that belief. Some virtue epistemologists talk about two levels of epistemic virtues. (Baehr, 2006; Lepock, 2014; Fairweather, 2014) We have some knowledge-generating processes that produce new information from perception, memory, or deduction. Then there is a higher-level type of epistemic virtues such as consciousness, humility, self-control, goodwill, originality, creativity. Baehr calls the latter 'good intellectual character traits.' (Baehr, 2006) There is a virtue epistemology primarily concerned with the first type of virtue (virtue reliabilists), while virtue responsibilists focus on the high level type. (Greco, 2000)

In the line of this virtue epistemology, Carter recently has analyzed trust as a bi-level concept and provided different definitions of trust. (Carter, 2019) He defines a type of trust called 'fully-apt trust'. First, there are some cognitive and externalist requirements on A. H hopes A will take care of things, and that A can successfully fulfill the actions when A is in a 'proper shape and properly situated.' But this are not the only conditions on A: H fully apt trust A when H's trust is 'convictive': "it is aptness on the first order guided by apt awareness on the second order that the first order performance would be apt (likely enough)." (Carter, 2019) 22 This third condition is metacognitive and improves the quality of H's trust in A when A is able to reflect through as a second order process upon its own error in thinking and upon its own confidence in reasoning.

We intend to adopt a naturalized virtue epistemology which is shaped by empirical evidence and to give a formal expression of this second-order monitoring competence suggested by responsibilists: the immediate option is to think in terms of metacognitive requirements.

---

the shelf.

2.b Metacognition and agency

Humans are often aware of errors in making decision and are able to report levels of confidence. The subject is often aware of the difficulty of making a decision. Cognitive science literature shows that these confidence levels are sometimes correlated with objective performance. Behavior is guided by these assessments of decision quality especially when there is no independent feedback available. In the case of AAA agents, when feedback or reward are not available, one can intuit the importance of self-assessment of performance.

The study of the disposition to reflect upon our own performance in memory, perception, learning, reasoning, communicating, and, ultimately, upon our own limit of knowledge, is an area of cognitive science called 'metacognition.' (Nelson, 1996; Proust, 2007; Fleming and Frith, 2014) As the processes of self-reflection, monitoring and controlling cognition, it is a superior one of the most sophisticated cognitive dispositions of mature humans and possibly a uniquely human cognitive disposition (psychologists still debate whether animals or infants have it).

In the heyday of the booming literature on metacognition, Nelson proposed some principles of metacognition. He suggested that in humans and animals, mental processes are divided into a low-level (object-level) and a high-level (meta-level). (Nelson, 1990) The high-level is always a model of the low-level and the two levels are related by causal relations such as: monitoring, controlling, and correcting. The control flow entails that the high-level causally influences the low-level by initiating, sustaining, or terminating activity.

As experimental philosophy and cognitive science have it, the ability to assess its own actions and knowledge is a superior competence. Together with Carter, we can use this definition in our approach to trust. But we need to see whether the agents can run an independent process of assessing their confidence, and this metacognitive requirement can be linked directly to the rationality of both A and H.

3 Rationality of trust and a plea for a Bayesian approach

3.a Rationality of trust

As a complex construct, trust involves the nature of both H and A, as well as A's future actions. As expected, a philosophical analysis of trust can be embedded in both epistemology and ethics. But for the sake of the present argument, we focus on its epistemology and just marginally on its ethics. The literature on trust refers frequently to its rationality: when do we have enough warrants to trust somebody (or something)?

Trust of H in A is rational when there are enough reasons to believe that A acts rationally in some substantial way. We advance here a thesis about trust: the requirement of A being rational can be couched in terms of probabilities, con-

ditionalizations, and computational processes. It is clear that a rational agent A is more trustful than a less rational agent A'. In the line of the naturalized approach preferred here, we relate rationality of trust to some metacognitive processes that A is able to instantiate. Trust and its relation to rationality have interesting epistemic implications. (Gauthier, 1987; Hieronymi, 2008; Faulkner, 2014; Rosenkranz, 2015) Our claim is that rationality of trust needs to include metacognitive components. This condition is, we argue, suitable to a case in which A is a human agent and when A is a technology, and especially an AI algorithm. Restrictive approaches to trust insist that 'genuine trust' has solid epistemological and ethical components. A mere description of what trust was historically or factually is not enough to ground trust in new technologies: we have not faced the challenges of AI, or synthetic life, or artificial life before. What means to be trustful when it comes to non-human agents?

We side here with a cognitive or rational approach to trust, in which trust is the product of a rational process. Trust in artificial agents implies a process of rational choice occurring in H and in A. The agent, be it artificial or not, has a set of goals, makes some choices and is able to calculate costs and gains of its own action. We argue that relating trust to cognitive and metacognitive abilities of the artificial agent has advantages over some non-cognitive approaches to trust.

3.b Confidence and trust

Here is again one requirement discussed by Carter: H can trust A if A is successfully reliable enough, only when A is in a proper shape and properly situated. To improve the quality of trust, Carter suggests adding a 'convictive' requirement on trust: H is guided in her trust in A by A's second-order assessment of risk in making a decision.

We relate 'risk assessment' from virtue epistemology to operationalizable quantities as measured in experimental philosophy or cognitive science. We think that confidence level as expressed in cognitive science literature is a good starting point for a metacognitive approach to trust. As a second order process in the brain, confidence is a belief about the validity of our thoughts, actions and performance. (Timmermans Bert et al., 2012; Meyniel et al., 2015; Fleming and Daw, 2017). Confidence is strongly related to awareness and consciousness and can be formalized in various ways. In neuroscience, Bayesian approaches to confidence are grounded in the assumption that uncertainty is coded in natural neural networks. We can extend this assumption to artificial neural networks and hence to artificial agents based on machine learning algorithms (see next section).

4 A probabilistic approach to trust

4.a Bayesian trust: accuracy *vs.* confidence

How do we relate trust to metacognition? We contrast here mere *accuracy of* representation as low level knowledge with *confidence in* one's representation as a high-level type of knowledge. As Karmiloff-Smith put it: the knowledge in the system is not the same as knowledge for the system.(Clark and Karmiloff-Smith, 1993)

At the beginning, without any evidence, the trustor H can have a certain degree of belief that agent A will represent the world as accurate as possible. But as Bayesian epistemology goes, this is only a prior probability that will be updated based on the actions of A. In the line of our previous discussion on metacognition, let us suppose that A is able to update its own confidence level in a similar, Bayesian way.

Let us imagine now a simplified agent A that is designed to produce an accurate observation of a current state of the world. The agent builds a model of its environment and makes a decision about acting upon this environment. The world (the world is here the totally of the variables of the environment in which A perceives and acts) is in a state $w \in W$. To simplify, we can imagine that the world has only two possible states $W = (W_1,\ W_2)$ (in experiments discussed in cognitive science, W is just an object that is oriented in two directions: '1=left' or '2=right'). By convenience we can think of W_1 as represented by value '-1' and W_2 as represented by value '1.'[3] The stimulus that A receives is noisy and inaccurate. The agent makes one observation o, which is not identical to w, but it is roughly correlated with w, given some noise factor and distortion of perception. Unlike w which is discrete, o takes any value on a continuous spectrum, which can be normalized to the interval $o \in O = [-1; 1]$. This observation creates a change in the internal decision variable $\boldsymbol{X_{act}}$ which can follow a Gaussian distribution conditional on the world w and on observation o:

$$X_{act} \sim N(w, \sigma^2_{act})$$

In this first approximation, the agent A consists of a model of the world M with a set of decision variable $\boldsymbol{X_{act}}$, and a set of parameters θ. The output of this model is an action a (belonging to a binary set, $A = (A_1, A_2)$).

H then is trusting A as accuracy of representation and this is not a higher epistemic competency, in the line of Sosa's suggestion. A simply tries to replicate the state w of the world. The observation o at any moment will affects the decision variables X_{act} which will create action a.

In this simple approach, there is no metacognitive requirement on trust. The trust of H is defined as a conditional probability that A, as a model-building process, will represent *accurately* the world. We can simply define trust as a degree of belief that H has about the accuracy of A's representation of W:

[3] See most of the literature on "confidence" and "accuracy" in perception: (Meyniel et al., 2015; Fleming and Daw, 2017)

[1] *Accuracy-Trust: H*'s trust in A is rational when the probability that A's representation of the world a is 'accurate enough'—to a certain threshold T_{acc}: $t_{acc} = P\left(a = w, a, M, \theta, X_{act}\right) \geq T_{acc}$

We want to augment this cognitive requirement of trust with a metacognitive one. One key concept in metacognition is 'confidence.' As Fleming and Dew and others suggest, we have some evidence that in the human brain there is a second process of inference that builds another output, call it the confidence. (Meyniel et al., 2015; Fleming and Daw, 2017) In the line of the computational theory of the mind, we can consider 'confidence' as a computational process that happens in the brain, but has its own internal variable, call it X_{conf}. We expect a covariance between the internal variables X_{act} and X_{conf}. Fleming and Daw simulated this by a bivariate Gaussian distribution:

$$\begin{bmatrix} \mathbf{X_{act}} \\ \mathbf{X_{conf}} \end{bmatrix} = N\left(w, \Sigma\right);\ \Sigma = \begin{pmatrix} \sigma_{act}^2 & \rho\sigma_{conf}\sigma_{act} \\ \rho\sigma_{conf}\sigma_{act} & \sigma_{conf}^2 \end{pmatrix}$$

Without entering the technical details, we can assume that the confidence computational process is coupled, up to a certain coupling constant ρ , to the process that decides the action variable *a*. This second process outputs a parameter representing its own performance. In the framework adopted here, confidence and uncertainty are quantified as degrees of beliefs, and are both interpreted as Bayesian probabilities. (Meyniel et al., 2015) Natural neural networks can output such confidence probabilities. In its very metacognitive approach, confidence and uncertainty can be 'statistical summaries' of the decision-making process happening in a human brain or in an artificial neural network. (Meyniel et al., 2015)

The confidence is then represented as a conditional probability of obtaining the right inference a=w, given both the action model and the confidence model, a second order model)

$$t_{conf} = P\left(a = w | a, M, \theta, \begin{bmatrix} \mathbf{X_{act}} \\ \mathbf{X_{conf}} \end{bmatrix}\right)$$

We then propose a metacognitive definition of trust. H trusts agent A when the confidence output improves the degrees of belief of H compared to mere accuracy:

[2] *Confidence-Trust: H's* trust in A is rational when A's confidence level improves the accuracy of its representation of the world (and its actions): $t_{acc} \leq t_{conf}$ or:

$$P\left(a = w, a, M, \theta, X_{act}\right) \leq P\left(a = w | a, M, \theta, \begin{bmatrix} \mathbf{X_{act}} \\ \mathbf{X_{conf}} \end{bmatrix}\right)$$

5 Trust in autonomous artificial agents (AAA): requirements for model and data uncertainty

Some artificial agents based on Machine Learning algorithms and natural neural networks can implement confidence level mechanisms. But most neural networks are prone to overfitting. We can suspect here, without offering a full argument, that data uncertainty and model uncertainty can be on pair with metacognitive requirements discussed for human agents A. It is a possible start to impose some metacognitive requirements (in nature) on algorithms such that H is rational.

Model uncertainty is a very general term that refers to the ability of an algorithm to correct itself. Data uncertainty refers to its ability to select reliable data and discard data that is erroneous or noisy. Together with most of the literature on machine learning, we acknowledge that 'model uncertainty' and 'data uncertainty' are not present in standard, feed-forward neural network (Murphy, 2012; Kendall and Gal, 2017) Even the infamous Deep Neural Network approach does not instantiate, as far as it is usually interpreted, the metacognitive requirement *per se*. Even those networks which avoid overfitting and implement partially a data uncertainty, do not fare well in respect of model uncertainty. We need to model within ML networks their own limitations, such that they are more 'rational' in making decisions. When a rational agent has low confidence level in its own output, let us say in a classification problem, it needs to output a result of the form "unable to classify." In more technical terms, we need to take the predictive variance of a model more serious when we evaluate its output. This is a way to show that although data may be reliable and trustful, the model is not able to represent it as an output.

A new approach to neural network proposed recently suggests that a "confidence readout" can be associated to the dropout method of avoiding overfitting of neural networks. (MacKay, 2003; Srivastava et al., 2014; Gal and Ghahramani, 2016) Similarly, if one adopts the dual route model suggested before, one can identify brain-scale circuits and neural codes for uncertainty in the brain. (Meyniel et al., 2015) When we start with a random or very unlikely a priori model of the world, the dropout factor gives us an indicator of the correction path taken to improve the model. The confidence level of this type of AAA agent is then a combination of a data uncertainty and model uncertainty. In this case, we can be optimistic that processes happening in the brain that generate confidence level or statistic summaries can be implements and programmed in artificial neural networks. Overall, an agent with metacognitive abilities is able to monitor the decision-making process and becomes more trustworthy than an agent that lacks some processes. The metacognition literature emphasizes that this self-monitoring process improves accuracy and the overall predictive performance of the agent. There is nevertheless something more than performance when it comes to metacognition: for low level of confidence, the agent may not produce a prediction at all and would suspend its own judgment Confidence and error-correction are different processes in the mind, but are they different than the process of decision making? Fleming and Dew model confidence as a

second-order computation, running in parallel with the decision-making process.

[3] Artificial-*Trust (A-trust): H's* trust in an artificial autonomous agent (*AAA)* is rational when *AAA* implements (by design) a certain number of high-level mechanisms such as: model uncertainty, data uncertainty, and mechanisms to avoid overfitting.

We can envisage here a future research project in the metacognitive definition of trust by adding a requirement on *AAA* which is ethical in nature. Decision theory has been developed mostly for human, rational agents. Whether it can be adapted to accommodate artificial intelligence is one of the paths explored in the philosophical literature on decision theory. We are interested in a more general foundational issue: integrate ethics in decision theory (Colyvan et al., 2010), with a special emphasis on artificial agents. The A-trust requirement and the need for a metacognitive mechanism in AAA agents can be reframed in terms of multi-objective agents. One possible multi-objective algorithm is the one in which a third mechanism, with its internal variable $\boldsymbol{X}_{\boldsymbol{eth}}$ calculates a level of moral confidence in the action of the agent. In the line of an 'is-ought' division of judgments, one can take the internal decision variables $\boldsymbol{X}_{\boldsymbol{eth}}$ and $\boldsymbol{X}_{\boldsymbol{act}}$ as not being correlated, unlike the correlation between $\boldsymbol{X}_{\boldsymbol{conf}}$ and $\boldsymbol{X}_{\boldsymbol{act}}$. Multi-objective agents are discussed frequently in the literature. (Doumpos, 2013) The multi-objective agent takes all the variables and optimizes two functions, called here the factual action determined by $\boldsymbol{X}_{\boldsymbol{act}}$ and the normative action with another output (independent of both a and confidence). H then would trust agent AAA when the accuracy of its action is "better" morally and factually than the accuracy without these high-level mechanisms:

[4] *Moral-Confidence-Trust: H's* trust in *AAA* is rational when *AAA's* confidence level and the normative assessment improve the accuracy of its action in the world:

$$P\left(a = w, a, M, \theta, X_{act}\right) \leq P\left(a = w | a, M, \theta, \mathbf{X_{eth}}, \begin{bmatrix} \mathbf{X_{act}} \\ \mathbf{X_{conf}} \end{bmatrix}\right)$$

6 Conclusion

Trust can be framed, in the approach presented here, as a set of metacognitive requirements. In its simplest form, we suggest that confidence level improves trust when the conditionalized probability with confidence is greater than the probability due to accuracy. We find important analogies between the way confidence is processed in the natural neural networks and artificial networks. This suggests that a metacognitive requirement on trust can be imposed on artificial agents.

References

Baehr, J. (2006, April). Character, reliability and virtue epistemology. *The Philosophical Quarterly 56*(223), 193–212.

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

Carter, J. A. (2019). On behalf of a bi-level account of trust. *Philosophical Studies*, 22.

Clark, A. and A. Karmiloff-Smith (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language 8*(4), 487–519.

Colyvan, M., D. Cox, and K. Steele (2010, September). Modelling the moral dimension of decisions. *Nous 44*(3), 503–529.

Cook, K. (Ed.) (2003, March). *Trust in Society*. New York, NY: Russell Sage Foundation.

Doumpos, M. (2013). *Multicriteria decision aid and artificial intelligence links, theory and applications*. Hoboken, NJ: Wiley-Blackwell.

Fairweather, A. (2014, May). *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*. Springer. Google-Books-ID: xJgpBAAAQBAJ.

Faulkner, P. (2014). The practical rationality of trust. *Synthese 191*(9).

Fleming, S. M. and N. D. Daw (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review 124*(1), 91–114.

Fleming, S. M. and C. D. Frith (Eds.) (2014). *The cognitive neuroscience of metacognition*. Heidelberg: Springer. OCLC: ocn879416801.

Gal, Y. and Z. Ghahramani (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33 rd International Conference on Machine Learning 48*, 10.

Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations*, Volume 13, pp. 213–237. Oxford: Basil Blackwell.

Gauthier, D. (1987, May). *Morals by Agreement*. Oxford University Press.

Greco, J. (2000). Two kinds of intellectual virtue. *Philosophy and Phenomenological Research 60*(1), 179–184.

Hawley, K. (2014). Trust, distrust and commitment. *Noûs 48*(1), 1–20.

Hieronymi, P. (2008, June). The reasons of trust. *Australasian Journal of Philosophy 86*(2), 213–236.

Hollis, M. (1998, March). *Trust Within Reason*. Cambridge University Press. Google-Books-ID: 8Ip0fjRa5VAC.

Hosking, G. A. (2014). *Trust: a history*. Oxford: Oxford University Press.

Jones, K. (1996). Trust as an affective attitude. *Ethics 107*(1), 4–25.

Kendall, A. and Y. Gal (2017, March). What uncertainties do we need in bayesian deep learning for computer vision? *arXiv:1703.04977 [cs]*. arXiv: 1703.04977.

Lepock, C. (2014). Metacognition and intellectual virtue. In A. Fairweather (Ed.), *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*, Synthese Library, pp. 33–48. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-04672-3_3.

MacKay, D. J. C. (2003, October). *Information Theory, Inference and Learning Algorithms* (1 edition ed.). Cambridge, UK ; New York: Cambridge University Press.

Meyniel, F., M. Sigman, and Z. Mainen (2015, October). Confidence as bayesian probability: From neural origins to behavior. *Neuron 88*(1), 78–92.

Murphy, K. P. (2012). *Machine learning, a probabilistic perspective*. Cambridge, Mass.: MIT Press.

Nelson, T. O. (1990, January). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of Learning and Motivation*, Volume 26, pp. 125–173. Academic Press. DOI: 10.1016/S0079-7421(08)60053-5.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist 51*(2), 102–116.

Nickel, P. J. (2017, March). Being pragmatic about trust. In P. Faulkner and T. Simpson (Eds.), *The Philosophy of Trust*. Oxford University Press.

Proust, J. (2007, November). Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese 159*(2), 271–295.

Rosenkranz, S. (2015, September). Fallibility and trust. *Noûs 49*(3), 616–641.

Sosa, E. (2007, June). *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford University Press.

Sosa, E. (2018, November). *Virtue Epistemology and a Theory of Competence*. Princeton University Press.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 1929–1958.

Timmermans Bert, Schilbach Leonhard, Pasquali Antoine, and Cleeremans Axel (2012, May). Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1594), 1412–1423.