Trustworthiness of science in the nexus between science, society and policy

Sonora, Marina

MARINA SONORA **2020** **DOKTORSKI RAD**

University of Zagreb

Faculty of Humanities and Social Sciences

Marina Sonora

TRUSTWORTHINESS OF SCIENCE IN THE NEXUS BETWEEN SCIENCE, SOCIETY AND POLICY

DOCTORAL THESIS

Supervisor: Dr. Hrvoje Jurić

Zagreb, 2020

Sveučilište u Zagrebu

Filozofski fakultet

Marina Sonora

VJERODOSTOJNOST ZNANOSTI U MEĐUODNOSU IZMEĐU ZNANOSTI, DUŠTVA I POLITIKE

DOKTORSKI RAD

Mentor: izv. prof. dr. sc. Hrvoje Jurić

Zagreb, 2020.

Mentor

Hrvoje Jurić is an associate professor at the Department for Philosophy at the Faculty for Humanities and Social Sciences, University of Zagreb. As a researcher and lead researcher, he participated in the implementation of seven scientific-research projects funded by the Ministry of Science, Education and Sports of the Republic of Croatia and the University of Zagreb as well as in four international projects funded by DAAD, Volkswagen-Stiftung, UNESCO and the European Commission.

He presented at more than a hundred scientific conferences both internationally and in Croatia, and participated in about twenty scientific and professional conferences as a listener and discussant. In addition, he held seventy public lectures on philosophical, scientific and socio-political topics in Croatia and abroad and presented at numerous public forums and book promotions. He is deputy editor-in-chief of the publications *Filozofska istraživanja* and *Synthesis Philosophica* and the editions *Filozofska istraživanja* at the Croatian Philosophical Society. He is also a member of the editorial board of the scientific journals *Jahr, The Holistic Approach to Environment* and *In medias res* and a member of the advisory board of the journal *Časopis za primijenjene zdravstvene znanosti*. He has edited over thirty publications for domestic and international scientific conferences, as well as several other publications.

Books: *Filozofija i rod* (zbornik radova), ur. Gordana Bosanac, Hrvoje Jurić, Jasenka Kodrnja, Hrvatsko filozofsko društvo, Zagreb, 2005.; *Etika odgovornosti Hansa Jonasa*, Pergamena, Zagreb, 2010.; *Filozofija i mediji* [zbornik radova], ur. Hrvoje Jurić, Sead Alić, Hrvatsko filozofsko društvo – Centar za filozofiju medija i mediološka istraživanja, Zagreb, 2014

Selected Articles: Princip očuvanja života i problem odgovornosti, *Filozofska istraživanja*, 71 (4/1998), str. 895-900; također u: Ante Čović (ur.), *Izazovi bioetike*, Pergamena – Hrvatsko filozofsko društvo, Zagreb, 2000., str. 141-148; (Ne)mogućnost jedne sveobuhvatne poetike, u: Ljerka Schiffler (ur.), *Zbornik o Frani Petriću*, Hrvatsko filozofsko društvo, Zagreb, 1999., str. 405-413.; Filozofijska hermeneutika i praktična filozofija Hans-Georga Gadamera, *Filozofska istraživanja*, 79 (4/2000), str. 615-650.; Najranija recepcija Heideggerove filozofije kod nas: Vladimir fra Kruno Pandžić, u: Pavo Barišić (ur.), *Otvorena pitanja povijesti hrvatske filozofije*, Institut za filozofiju, Zagreb, 2000., str. 377-395.; Priroda i etika, u: *Sedra rijeke Une i Una bez sedre*, Privredna komora Unsko-sanskog kantona – Društvo Unski smaragdi, Bihać, 2000., str. 101-112.; Gadamer o odnosu teorije i prakse u filozofiji i medicini, u: Damir Barbarić,

Tomislav Bracanović (ur.), *Gadamer i filozofijska hermeneutika*, Matica hrvatska, Zagreb, 2001., str. 145-176.; Mora li tubitak jesti? Heideggerova analiza tubitka i Jonasova filozofijska biologija, *Filozofska istraživanja*, 84 (1/2002), str. 37-53.; Etika i politika, bioetika i biopolitika, *Socijalna ekologija*, god. 11 (2002), sv. 3, str. 233-244.; Rađanje s predumišljajem. Ili: Sjećaš li se Louise Brown?, *Treća*, god. 4 (2002), sv. 2, str. 128-151.; Tko i što je čovjek? Napomene uz povijest ograničavanja pojma 'čovjek', *Filozofska istraživanja*, 90 (3/2003), str. 753-760.; Utopie – Anti-Utopie – Post-Utopie – Utopie. Zu Jonas' Kritik des marxistischen Utopismus, *Synthesis philosophica*, 35-36 (1-2/2003), str. 207-225; također: Utopija – anti-utopija – post-utopija – utopija. Uz Jonasovu kritiku marksističkog utopizma, *Filozofska istraživanja*, 91 (4/2003), str. 1141-1156.; Philosophische Zeitschriften in Kroatien von den 90er Jahren des 20. Jahrhunderts bis heute, u: Georgi Kapriev (ur.), *Philosophie in Südosteuropa. Stand der Forschung und der Veröffentlichungen*, Iztok-Zapad, Sofia, 2004., str. 65-78.; Kangrgina riječ o zavičaju, *Filozofska istraživanja*, 94-95 (3-4/2004), str. 757-762.; Svijet kao samovolja i predrasuda. Schopenhauer o spolnosti i o ženama, *Filozofska istraživanja*, 99 (4/2005), str. 791-804; također u: Gordana Bosanac, Hrvoje Jurić, Jasenka Kodrnja (ur.), *Filozofija i rod*, Hrvatsko filozofsko društvo, Zagreb, 2005., str. 49-65.; Utjecaj Katoličke crkve u Hrvatskoj na politiku reproduktivnih i seksualnih prava i zdravlja, u: Simona Goldstein (ur.), *Otvorenost društva – Hrvatska 2005.*, Institut Otvoreno društvo, Zagreb, 2005., str. 166-198 (s Marinom Škrabalo).; Utopija – anti-utopija – post-utopija – utopija. Ili: utopija, filozofija i društveni život, *Arhe*, 1/2005, str. 217-225.; Rasprava o knjizi *Utopija i inauguralni paradoks* Gordane Bosanac, *Filozofska istraživanja*, 101 (1/2006), str. 137-149 (s Gordanom Bosanac, Linom Veljakom i Marijanom Krivakom).; Zdravlje i vrijednosti, *Hrvatski časopis za javno zdravstvo*, god. 2 (2006), sv. 6; također u: Ana Borovečki, Slobodan Lang (ur.), *Javno zdravstvo, etika i ljudska prava*, Sveučilište u Zagrebu, Medicinski fakultet, Škola narodnog zdravlja "Andrija Štampar", Zagreb, 2010., str. 104-108.; Žene i priroda. Prilozi za kritiku ekofeminizma, u: Ankica Čakardić, Ana Jelušić, Daniela Majić, Tanja Ratković (ur.), Kategorički feminizam. Nužnost feminističke teorije i prakse, Centar za ženske studije, Zagreb, 2007., str. 97-110.; Contemporary Croatian Philosophy, u: Maija Kule (ur.), *Philosophy Worldwide: The Current Situation. Materials for the International Cooperation and Philosophical Encounters*, FISP – University of Latvia, Riga, 2007. (ISBN: 978-9984-624-51-8), str. 102-115 (s Mislavom Kukočem).; Stützpunkte für eine integrative Bioethik im Werk Van Rensselaer Potters, u: Ante Čović, Thomas Sören Hoffmann

(ur.), *Integrative Bioethik / Integrative Bioethics*, Academia Verlag, Sankt Augustin, 2007., str. 68-92; također: Uporišta za integrativnu bioetiku u djelu Van Rensselaera Pottera, u: Velimir Valjan (ur.), *Integrativna bioetika i izazovi suvremene civilizacije*, Bioetičko društvo u BiH, Sarajevo, 2007., str. 77-99; također u: Ana Borovečki, Slobodan Lang (ur.), *Javno zdravstvo, etika i ljudska prava*, Sveučilište u Zagrebu, Medicinski fakultet, Škola narodnog zdravlja "Andrija Štampar", Zagreb, 2010., str. 33-53; također: Potpirnite točki za integrativnata bioetika vo deloto na Van Renselar Poter, *Filozofija*, 31/2011, str. 19-31.; Uz *Život životinjâ* J. M. Coetzeeja i replike na nj, u: Suzana Marjanić, Antonija Zaradija-Kiš (ur.), *Kulturni bestijarij*, Institut za etnologiju i folkloristiku – Hrvatska sveučilišna naklada, Zagreb, 2007. (ISBN: 978-953-6020-36-2), str. 485-492.; Bioetika u Hrvatskoj, *Filozofska istraživanja*, 111 (3/2008), str. 601-611 (s Ivanom Zagorac).; Humanizam, transhumanizam i problem rodno-spolne drugosti, *Treća*, god. 10 (2008), sv. 1, str. 27-44.; Ugrožavanje prirode i kulture kao izazov za bioetiku i multikulturalizam, *Philological Studies*, god. 6 (2008), sv. 1; također u: Velimir Valjan (ur.), *Integrativna bioetika i interkulturalnost*, Bioetičko društvo u BiH, Sarajevo, 2009., str. 83-92.; Životinjska duša i životinjska prava. Pitanja i odgovori (o) filozofiji Hansa Jonasa, *Arhe*, br. 12/2009, str. 107-120; također: Tierseele und Tierrechte. Fragen und Antworten zur Philosophie Hans Jonas', u: Walter Schweidler (ur.), *Wert und Würde der nichtmenschlichen Kreatur / Value and Dignity of the Nonhuman Creature*, Academia Verlag, Sankt Augustin, 2010., str. 111-123.; Euthanasia in the Context of Croatian Healthcare System, Legislation and Bioethical Discussions, u: Brigitte E. S. Jansen, Nada Gosić (ur.), *Croatia: Politics, Legislation, Patient's Rights and Euthanasia*, Martin Meidenbauer Verlagsbuchhandlung, München, 2011., str. 93-117.; O djelu i liku filozofa i profesora Ante Pažanina, *Filozofska istraživanja*, 123 (3/2011), str. 491-498.; Odgovornost za dijete kao paradigma bioetičke odgovornosti, u: Ante Čović, Marija Radonić (ur.), *Bioetika i dijete. Moralne dileme u pedijatriji*, Pergamena – Hrvatsko društvo za preventivnu i socijalnu pedijatriju, Zagreb, 2011., str. 49-62.; Crucifixion of the Identity: Persons and Beings, Bodies and Genes, u: Sibila Petlevski, Goran Pavlić (ur.), *Spaces of Identity in the Performing Sphere*, Fraktura – Akademija dramske umjetnosti, Zagreb, 2011., str. 163-187.; Feminism in the Light of the Bioethical Pluri-Perspectivism, u: Ante Čović (ur.), *Integrative Bioethik und Pluriperspektivismus / Integrative Bioethics and Pluri-Perspectivism*, Academia Verlag, Sankt Augustin, 2011., str. 237-243.; Hans Jonas' Integrative Philosophy of Life as a Foothold for Integrative Bioethics, *Jahr*, god. 2 (2011), sv. 4, str. 511-520; također u: Amir

Muzur, Hans-Martin Sass (ur.): *Fritz Jahr and the Foundations of Global Bioethics: The Future of Integrative Bioethics*, Münster – Berlin – Wien: Lit Verlag, 2012., str. 139-148.; Multi-Disciplinarity, Pluri-Perspectivity and Integrativity in the Science and the Education, *The Holistic Approach to Environment*, god. 2 (2012), sv. 2, str. 85-90.; Odjeća, tijelo bez odjeće, golo tijelo, tijelo. Golotinja u suvremenoj umjetnosti, u: Irfan Hošić (ur.), *Odjeća kao simbol identiteta*, Univerzitet u Bihaću, Tehnički fakultet, Bihać, 2012., str. 99-115.; Privatisation of Life, u: Walter Schweidler (ur.), *Bioethik – Medizin – Politik / Bioethics – Medicine – Politics*, Academia Verlag, Sankt Augustin, 2012., str. 97-104.; Bioetički aspekti upravljanja vodnim dobrima, u: Mile Beslić, Dario Ban (ur.), *Aktualna problematika u vodoopskrbi i odvodnji*, Revelin, Ičići, 2012., str. 497-504 (s Tomislavom Krznarom).; Ekofeminizam vs. ekologija za žene, u: Rada Drezgić, Daša Duhaček, Jelena Vasiljević (ur.): *Ekofeminizam: nova politička odgovornost*, Institut za filozofiju i društvenu teoriju, Beograd, 2012., str. 42-61.; Anarhizam i marksizam u perspektivi "praxis-filozofije", u: Dragomir Olujić Oluja, Krunoslav Stojaković (ur.), *Praxis – društvena kritika i humanistički socijalizam*, Rosa Luxemburg Stiftung, Beograd, 2012., str. 173-200.; Od Hirošime do Fukušime: nuklearna tehnologija nekoć i danas, u: Velimir Valjan (ur.), *Integrativna bioetika pred izazovima biotehnologije*, Bioetičko društvo u BiH, Sarajevo, 2012., str. 19-31.; Whitmanovo proročanstvo. O poeziji, medijima i demokraciji, u: Hrvoje Jurić, Sead Alić (ur.), *Filozofija i mediji*, Hrvatsko filozofsko društvo – Centar za filozofiju medija i mediološka istraživanja, Zagreb, 2014., str. 395-406.; Scientific De(con)struction and Artistic (Re)construction of the Body, u: Claudia Bosse (ur.), *Struggling Bodies in Capitalist Societies (Democracies)*, Vienna: Cheap Method Edition, 2014., str. 74-80.; Samoupravljanje – prije i poslije socijalizma, u: Lino Veljak (ur.), *Gajo Petrović, filozof iz Karlovca*, Hrvatsko filozofsko društvo, Zagreb, 2014., str. 101-112.; Krležina "Evropa danas" u Europi danas, u: Boris Gunjević (ur.), *Krleža za ponavljače*, Sandorf, Zagreb, 2014., str. 82-95.; Život usred života: zašto i kako je nastajala bioetika, *Sarajevske sveske*, 47-48/2015, str. 13-23.; From the Notion of Life to an Ethics of Life, *Synthesis philosophica*, 59 (1/2015), str. 33-46.

Acknowledgment

This dissertation is dedicated to my husband, Dr. Robert J. Sonora who inspires the best in me and encouraged me to finalize my dissertation. I also dedicate this dissertation to my parents for teaching me how to live life with grit, hard work and resilience, and to my sisters, family, and friends for always being there for me.

I would like to especially acknowledge the Wirth Institute for Austrian and Central European Studies at the University of Alberta, Canada, directed by Dr. Joseph Patrouch, for awarding me Doctoral Fellowship that allowed me to conduct my research, present and attend conferences in Canada and USA. I am grateful for insightful comments to my work on the dissertation presented at the University of Alberta, at Ryerson University in Toronto, Canada, and at the University of British Columbia in Vancouver, Canada.

I would like to thank my advisor and the rest of my thesis committee for their insightful comments and encouragement. My sincere thanks also goes to my colleagues at the Department for Horizon 2020, Research and Innovation Program, Agency for Mobility and EU Progarmmes, and my project partners who inspired me to widen my research from various perspectives. Finally, I would like to express my gratitude to ATG/Cognizant for support and continuous opportunity for growth.

Abstract

Although trusting relations have been attributed as crucial for the social fabric, trust is in flux. Erosion of trust is increasingly being captured in surveys, recent studies, and media accompanied by multiple concepts of trust envisaging ways of restoring it. The main aim of this thesis is to reframe the problem from restoring trust to the question of trustworthiness. On that ground, we defend the thesis that the trustworthy by design model is better suited to address the question of distrust and misplaced trust than the alternatives. Based on the critical analysis of the state-of-the-art concepts of trust and trustworthiness we propose a new model of *trustworthy by design* that accommodates more inclusive aspects of responsibility, responsiveness and the diversity of perspectives. In order to test the model, we apply it to two case studies.

First, we apply the trustworthy by design model to socio-technological algorithmic systems in decision-making process. It incorporates both responsibility and responsiveness early in the process of algorithmic design and throughout the decision-making cycle from framing the problem, choosing the data and the model, and all the way to deployment. We identify the input from diverse actors, domain experts, fairness, and inclusion as critical ingredients when making tradeoffs throughout the decision-making process. Second, we address the challenge of responsibility in distributed systems, governing challenges, regulation, and accountability.

Finally, we apply the trustworthy by design model in the context of scientific institutions. Within the broader context of science and values we identify main indicators that influence trust in science from framing the problem and the scientific agenda to diverse voices, confirmation, and macro-biases. By applying the trustworthy by design model we argue that trustworthy scientific institutions require reformulation of the notion of objectivity in science to ensure responsibility, responsiveness and democratic accountability.

Keywords:
trustworthy by design, socio-technological algorithmic systems, scientific institutions, responsibility, responsiveness, accountability

Proširenisažetak

Povjerenje je nužno u našim odnosima s drugim ljudima, javnim institucijama, organizacijama te digitalnim platformama. Povjerenje može biti terapeutsko (Horsburgh, 1960), korisno za suradnju (Gambetta, 1988; McLeod, 2015), samopoštovanje (Govier, 1993; McLeod, 2015) ekonomski i društveni prosperitet (Fukuyama, 1996), itd. Pa ipak, gubitak povjerenja trend je koji bilježe istraživanja javnog mnijenja, studije i mediji.

Istovremeno, promjene u okviru digitalnih transformacija, koje neki zovu četvrtom industrijskom revolucijom (Schwab, 2017), usmjeravaju naše povjerenje i nepovjerenje prema novim digitalnim tehnologijama. Istraživanja te problematike ukazuju na promjenu modusa povjerenja u distribuirano povjerenje putem tehnologija i povjerenja u platforme (Botsman, 2017), istraživanje našeg povjerenja u tehnologije i artificijelne aktere (Taddeo, 2010; Castelfranchi and Falcone, 2010; Buechner and Tavani, 2011) te povjerenja u algoritme (Cohen et al., 2014; Rubel and Jones, 2016; Shackelford and Raymond, 2014).

Najčešći odgovori na gubitak povjerenja bazirani na istraživanjima javnog mijenja te višestrukim konceptima povjerenja imaju za cilj pokušaj povratka povjerenja ili pak posve oprečno smatraju da je povjerenje nepotrebno u institucijama te da povjerenje u institucije nije primjenjivo u kontekstu novih digitalnih tehnologija.

Kritičkom analizom trenutnih dosega istraživanja najznačajnijih koncepata povjerenja (Baier, 1986; Jones, 1996, 2013, 2017; Dasgupta, 1988; Cook et al., 2005; Hardin, 2002, 2006; O'Neill, 2002a, 2014; Townley and Garfield, 2013) pratimo promjene iz odnosa povjerenja između pojedinaca, do povjerenja u institucije te transformacije povjerenja u digitalne tehnologije. Ta pluridimenzionalna mapa koncepata povjerenja i njihovih jedinstvenih obilježja vezanih uz kontekst primjene služi kao prvi korak u adresiranju pitanja povjerenja u druge, u institucije, digitalne platforme ili algoritmičke sustave te određivanja emocionalnih ili racionalnih karakteristika povjerenja.

Temeljem rezultata analize postavljamo tri međusobno povezane hipoteze. Prvo, ukazujemo na važnost modusa povjerenja (tzv. *strong thin mode*) koji ima veliki utjecaj na naše živote, primjerice na zdravlje ili financijsku stabilnost, ali je istovremeno naše razumijevanje povjerenja ograničeno jer se odvija u kompleksnim organizacijskim uvjetima gdje direktan kontakt licem u lice nije dostupan (Hosking, 2014). Ovaj modalitet povjerenja, iako zapostavljen,

ukazuje na polugu koja nedostaje u razumijevanju transformacije između različitih modusa povjerenja.

Drugo, na temelju razrade tvrdnje o neodrživosti hipoteze o deficitu znanja, tvrdimo da afektivni, emocionalni i kognitivni aspekti povjerenja igraju značajnu ulogu. Treće, na temelju kritičke analize različitih koncepata povjerenja, namjesto napuštanja povjerenja u institucije, predlažemo model koji bi bolje rješavao uočene nedostatke u modelu vjerodostojnih institucija.

Temeljem kritičke analize koncepata povjerenja i vjerodostojnosti, cilj je rada preformulirati pitanje povjerenja u pitanje o vjerodostojnosti te izložiti novi model vjerodostojnog dizajna koji obuhvaća aspekte odgovornosti i dijaloga. Na toj osnovi branimo tezu da je model vjerodostojnosti po dizajnu prikladniji za rješavanje pitanja nepovjerenja i pogrešnog povjerenja nego alternative.

Taj obrnuti pristup temeljen na vjerodostojnosti namjesto povjerenja nije nov, nekoliko autora je istaklo njegovu važnost u međuljudskim odnosima (Baier, 1986; Pettit, 1995; Jones, 1996, 2012; McGeer, 2008), te u institucionalnom i društvenom kontekstu (Hardin, 1996, 2002; O'Neill, 2002a, 2013, 2014; Potter, 2002). Nadalje, konceptualni okvir o razlici između koncepta povjerenja i vjerodostojnosti temeljimo na radu Onore O'Neill (2002a, 2002b, 2013, 2014) koja naglašava važnost davanja povjerenja jedino vjerodostojnim pojedincima ili institucijama (O'Neill, 2013). Na tom tragu dajemo prednost konceptu vjerodostojnosti nad povjerenjem s ciljem određivanja njegovih značajki koje mogu doprinijeti povjerenju u vjerodostojne institucije, sustave ili pojedince.

Teza koju predstavljamo o modelu pouzdanog dizajna institucija ili sustava razlikuje se od postojećih pristupa, modificirajući koncept vjerodostojnosti koji sadrži aspekte odgovornosti, dijaloga i raznolikosti perspektiva. Nadalje, u svrhu testiranja model primjenjujemo na dvije studije slučaja o algoritamskim socio-tehnološkim sustavima te znanstvenim institucijama ukazujući na važnost njegove primjene u međuodnosu znanosti, društva i politike u okviru 4. industrijske revolucije. Predloženo preoblikovanje sustava i institucija trebalo bi omogućiti dvostrano povjerenje između javnosti i sustava koji se temelje na raznolikosti i uključivosti.

U prvom poglavlju primjenjujemo fokus je na socio-tehnološkim algoritamskim sustavima koji su u pozadini digitalnih platformi. Posebno nas zanimaju algoritmi za donošenje odluka koji imaju veliki utjecaj na odluke o našem životu od zdravstvene skrbi, kreditnog skoringa do pravnog sustava. Mogu li algoritmi biti vjerodostojni, mogu li donositi pravedne

odluke? U distribuiranim sustavima umjetnih i ljudskih agenata tko treba biti odgovoran kada nešto pođe po zlu? Kako regulirati algoritme ako su neki od njih crne kutije čak i njihovim kreatorima?

Mapiranjem najutjecajnijih algoritama za donošenje odluka usredotočujemo se na pokazatelje koji utječu na gubitak povjerenja javnosti. To nas vodi pobližem ispitivanju utjecaja vrijednosnih aspekata algoritama i propusta u odgovornosti koji su posebno prisutni kod algoritama strojnog učenja, uz prijedlog načina osiguravanja vjerodostojnosti algoritamskih sustava.

Usprkos lakoći s kojom dajemo povjerenje digitalnim platformama detaljnija analiza infrastrukture u pozadini ukazuje na složeniju problematiku jer algoritmi nisu neutralni, a pitanje odgovornosti u kompleksnom suodnosu artificijelnih i ljudskih aktera daleko od je od jednoznačnog. Mapiranje indikatora koji utječu na promjene povjerenja provodimo na primjeru algoritama za donošenja odluka, među kojima su algoritmi personalizacije (Hildebrandt, 2008; Newell and Marabelli, 2015; Taddeo and Floridi, 2015), algoritmi za profiliranje (Hildebrandt, 2008), algoritmi strojnog učenja, (Tutt, 2016; Burrell, 2016), algoritmi pregovaranja (Raymond, 2015), itd.

Uz prednosti algoritamskih sustava u doprinosu učinkovitosti i poboljšanja sposobnosti, rezultati analize ukazuju na dva glavna izazova. Prvi se odnosi na vrijednosti prisutne u dizajnu algoritama, tehničkih ograničenja ili vrijednosti u procesu njihove provedbe. Algoritmi sadrže vrijednosne procjene koje je moguće identificirati u obliku rasnih predrasuda primjerice diskriminacije zbog boje kože (Noble, 2018), rasnih predrasuda u prediktivnom radu policije (Lum i Isaac, 2016, Ferguson, 2017, Angwin i sur., 2016), ili pak u kliničkim ispitivanjima (Kurt et al., 2016) itd.

Rodne predrasude, s druge strane, prisutne su u reklamama koje targetiraju muškarce više nego žene za bolje plaćene poslove (Datta i sur., 2015, Campolo i sur., 2017), ili za karijere u STEM-u (Lambrecht i Tucker, 2016). Također, ugrađivanje riječi koje se često koristi u algoritmima strojnog učenja za rangiranje pretraživanja Interneta (Nalisnick i sur., 2016) ili analiza životopisa (Hansen i sur., 2015) mogu ukazivati na predrasude (Bolukbasi i sur., 2016) kao što je primjerice kod riječi *muškarac* koja se češće pojavljuje uz riječ *računalni programer,* a riječ *žena* pored riječi *domaćica*. Na temelju analize algoritama identificirani pokazatelji koji

utječu na promjenu u povjerenju javnosti uključuju među ostalim predrasude, diskriminaciju te deficit po pitanju pravednosti, raznolikosti i uključenosti.

Analiza drugog izazova odgovornosti u distribuiranim sustavima ukazuje na izazove upravljanja, regulacije, odgovornosti, transparentnosti i angažmana javnosti kao pokazatelje koji utječu na promjene u povjerenju javnosti. Štoviše, koncept individualne odgovornosti gubi relevantnost u kontekstu kompleksnih sustava ljudskih i umjetnih aktera gdje je koncept distribuiranih agenata (Floridi 2013; Floridi i Taddeo, 2016) prikladniji za određivanje odgovornosti.

Na temelju analize identificiranih indikatora namjesto povjerenja u algoritme, vjerodostojnost treba razmatrati u kontekstu kompleksnog suodnosa između ljudskih i artificijelnih aktera u okviru socio-tehnoloških sustava. Algoritme nije moguće smatrati vjerodostojnima budući da ne posjeduju intencionalnost koja je svojstvena samo ljudima kao moralnim agentima. Posljedično, model vjerodostojnog dizajna odnosi se prvenstveno na socio-tehnološke algoritamske sustave u kojima su pitanja odgovornosti i dijaloga prisutna od ranog procesa dizajna.

U posljednja dva poglavlja model vjerodostojnosti primjenjujemo u kontekstu znanstvenih institucija u okviru šire rasprave o vrijednostima u znanosti u međuodnosu znanosti, društva i politika. Analiziramo trenutnu raspravu o idealu znanosti bez vrijednosti. S jedne strane zagovornici ideala bez vrijednosti u znanosti (McMullin, 1982; Lacey, 1999; Mitchell, 2004) zagovaraju isključivanje vrijednosti iz znanstvenog procesa. S druge strane, kritičari tvrde da postoji nekoliko mjesta za pozitivnu ulogu neepistemističkih vrijednosti u znanosti.

Prvi pristup temelji se na argumentu neodređenosti teorije (Harding, 1991; Longino, 1990, 2002) da je znanost nedovoljno određena dokazima što omogućuje prisutnost vrijednosti. Drugi je argument induktivnog rizika (Rudner, 1953; Churchman, 1948; Douglas, 2000, 2009), gdje znanstvenici suočeni s induktivnim rizikom moraju donositi vrijednosne sudove o tome koliko je dokaza dovoljno, ovisno o mogućim posljedicama lažno pozitivnih rezultata ili lažnih negativa. Treće mjesto za vrijednosti u znanosti je artikulirano (Kourany, 2010) unutar ideala društveno odgovorne znanosti koji uvodi društvene kao i epistemičke standarde.

Kontraverze povezane sa istraživanjima ukazuju da su uz koristi prisutni i potencijalni rizici uslijed epistemičke ovisnosti u znanosti. Zbog nedostatka sposobnost provjere svake znanstvene tvrdnje (kao što je Zemlja okrugla) često se moramo osloniti na znanje znanstvenih

stručnjaka i drugih (Hardwig, 1991). Stručnjaci mogu biti u neslaganju i znanstvene tvrdnje osporavane. Potencijalne pristranosti, te posljedice politizacije i komercijalizacije znanosti mogu dovesti u pitanje povjerenje u znanost. Analizom dvaju pristupa rješavanju navedenog izazova: (1) povjerenje u individualne stručnjake na temelju individualne stručne odgovornosti (Douglas, 2008, 2015) te (2) pristup institucionalnog nepovjerenja - utemeljen na institucionalno organiziranom skepticizmu (Merton, 1938; Bouchard, 2016) i transformativnoj kritici (Longino, 1990, 2002). Niti jedan od navedenih dvaju pristupa ne adresira širi kontekst društvenih uvjeta u kojima se provodi znanost.

U svrhu implementiranja konstitutivnih elemenata odgovornosti argumente nedovoljnog određivanja i induktivnog rizika stavljamo u kontekst u kojem znanstvene institucije djeluju u međuodnosu između znanosti, društva i politike, gdje dijalog i odgovornost prema javnosti igraju ključnu ulogu. Putem izvan-znanstvenih kritičkih doprinosa (Wylie, 2014) nadograđujmo koncept transformativnog kriticizma (Longino 1990, 2002) temeljenog na argumentu nedovoljnog određivanja. Drugo, na temelju argumenta induktivnog rizika (Douglas, 2000, 2009), tvrdimo da se demokratska odgovornost može postići u institucionalnom okruženju osiguravanjem otvorenosti o vrijednosnim sudovima te sudjelovanjem javnosti. Primjenom modela vjerodostojnog dizajna u institucijama i sustavima tvrdimo da vjerodostojne znanstvene institucije zahtijevaju preoblikovanje pojma objektivnosti u znanosti kako bi se osigurali dijalog i odgovornost.

Ključne riječi

povjerenje, vjerodostojnost, socio-tehnološki algoritamski sustavi, znanstvene institucije

Introduction

The benefits of trust and trusting relations are deeply embedded in society. Trust is necessary in our personal relationships with other people as well as in public institutions, organizations, and digital platforms. Trust can be therapeutic (Horsburgh, 1960), beneficial for cooperation (Gambetta, 1988; McLeod, 2015), self-respect (Govier, 1993; McLeod, 2015) economic and social prosperity (Fukuyama, 1996), etc. Decreasing trust would therefore rightly raise alarm because in a long run it can have serious societal consequences that relate to all of us.

And yet, erosion of trust is increasingly being captured in surveys, recent studies, and media. The results acquired by the Edelman Trust Barometer for 2017 (Edelman, 2017) conducted in 28 countries across the world showed the steepest decline in trust in four main institutions, government, business, media and, NGOs. Edelman Trust Barometer for 2018 (Edelman, 2018) shows no improvement of that trend. On the other side, the Eurobarometer Survey 2017 (European Commission, 2017) tracks 6% increase in optimism about the future of the EU, compared to 2016 including a majority of Europeans, 56% in sum.

Some of the recent examples that place trust under threat not only in institutions such as banks after the financial crisis but across a spectrum of institutions are detailed by Botsman:

> "The British MPs' expenses scandal; the false intelligence about weapons of mass destruction (WMDs); Tesco's horsemeat outrage; price gouging by big pharma; the BP Deepwater Horizon oil spill; the dishonours of FIFA's bribery; Volkswagen's 'dieselgate'; major data breaches from companies such as Sony, Yahoo! and Target; the Panama Papers and widespread tax avoidance; the exchange-rate manipulation by world's largest banks; Brazil's Petrobras oil scandal; the lack of an effective response to the refugee crises; and, last but not least, shocking revelations of widespread abuse by Catholic priests, other clergy and other 'care' institutions" (Botsman, 2017, p.4).

Botsman very vividly describes particular stories where loss of public confidence is caused by unethical behavior that have a huge impact on trust in institutions.

The debates on losing trust in institutions bypass neither the public trust in science and scientific institutions. Now well-known claim of Michael Gove, leave campaigner for Brexit that "people in this country have had enough of experts" (Mance, 2016) is somewhat misunderstood

as he was referring specifically to economic experts (Katz, 2017). Nevertheless, it has been taken up as one of the indicators of the current trend of decreasing trust in science and a possible extension of the distrust in other institutions. That indication has provoked debates within scientific discourse questioning if "science today (is) just the latest candidate for inclusion in the growing list of failing institutions that seems to characterize our society?" (Sarewitz, 2016, p. 38). Further suggestions about the causes of mistrust in science are numerous. Among others, they include the influence of fake news and social media echo chambers, groupthink, misconduct, research controversies, possible risks for society, exclusion of particular perspectives, preference, confirmation, and cognitive biases and potential unwarranted consequences of politicization or commercialization of science.

Simultaneously, rapid changes and digital transformations - within what some call the 4th industrial revolution (Schwab, 2017) - direct our trust and mistrust to new digital technologies. Some argue that trust is changing its form to distributed trust enabled through technology and trust in platforms (Botsman, 2017), another line of research examines trust in technology and artificial agents (Taddeo, 2010; Castelfranchi and Falcone, 2010; Buechner and Tavani, 2011) and trust in algorithms (Cohen et al., 2014; Rubel and Jones, 2016; Shackelford and Raymond, 2014).

The responses to decreasing trust are extensive, mostly based on surveys, multiple concepts of trust envisaging ways of restoring trust, to proposals that trust is not necessary in institutions or that institutional trust has not been designed for this digital era that we live in. However extensive and beneficial they are, there is one significant challenge that is mostly overlooked, namely they do not resolve the main challenge of trusting the untrustworthy.

As we have seen a prominent way of determining trust in institutions, including scientific institutions is by conducting surveys, although if we look closer to their results the decrease of trust in science is not so obvious. According to several surveys, trust in scientists or scientific institutions have ranked very high. According to Ipsos Mori's Veracity Index (Ipsos Mori, 2017), the percentage of people who trust scientists, to tell the truth, is very high 79%, according to the General Social Science Surveys (Smith and Son, 2013) the confidence in scientific institutions is very high. Eurobarometer Survey of the European Commission (Eurobarometer Special Survey, 2010) shows interest in science as high as 80%. According to those general results, it seems that

trust in science is not following the global trend of decline of trust in institutions such as government, business, media and NGOs (Edelman, 2017, 2018).

However, those generic measurements of trust in science as institutions can hardly be beneficial or useful in this general form (Hardin, 2006). They are far from giving an answer to the three-part formula that is mostly used in the research on trust, namely that A trusts B to do X (Hardin, 2002, p 9). Generic surveys about trust in science, therefore, do not clarify what is trusted, do we measure trust in science in general, in a scientific institution, which specific research field, which part of a scientific process, choice of a research problem, research methodology or research application. And who are the subjects of trust? If it is public it becomes unclear again which part of publics are the results about, because public is constituted of various groups. Finally, what is specific context and what is a valued thing that we measure trust for?

Apart from surveys, challenges of restoring trust could be identified in multiple concepts of trust in interpersonal and institutional context. Closer analysis of main current approaches to trust will expose various aspects that are inadequate in restoring trust in the trustworthy, from goodwill account of trust (Baier, 1986; Jones, 1996), trust as an affective attitude (Jones, 1996, 2013), encapsulated interest view (Hardin, 2002) to social norms approach (Dasgupta, 1988; Hardin, 2002; O'Neill, 2002a).

The third line of proposals provide profound insights into current transformations in trusting relationships. Be it that trust is not necessary in institutions since it can be replaced by other mechanisms (Cook et al., 2005), or that trust in institutions is not designed for digital age where distributed trust is predominant (Botsman, 2017). Although greatly significant, the transformations in trust - from individual relations to trust in institutions and to trust in digital platforms - should not mask the pertinent question that lies behind all of those forms of trust that exist simultaneously or successively, and that is the question of trusting the trustworthy. This challenge is not resolved in the digital era, but might be even more pertinent since it becomes so easy to trust the technology that the question on the trustworthiness of the algorithmic systems that run in the background can be even harder to challenge.

Next to decreasing trust, we sometimes also wrongly trust in the untrustworthy or distrust the one that is indeed trustworthy. It should raise a concern as "misplaced trust is pernicious, leads to disillusionment and embitterment, then, if it persists, to cumulative distrust, and ultimately to social breakdown. It is a destructive, not a constructive, force" (Hosking, 2014, p. 4,

chapter 3). Both misplaced trust as well as distrust, be it in other people, institutions or digital platforms raise a concern but often leads to initiatives for restoring trust. The prospect of restoring trust, however, is misplaced because rebuilding trust in untrustworthy leads only to repeating misplacing trust. It has a cumulative effect on distrust and causes repeating the vicious cycle of distrust. The question that should be asked instead is the one about trustworthiness, are we placing our trust in the one who is worthy of our trust.

The aim of this thesis, therefore, is to reframe the problem as the one about trustworthiness. That reframing then serves as a starting point for defending the thesis that both questions of trust and trustworthiness would be better addressed through trustworthiness by design. The trustworthy by design model will be illustrated through two case studies one in algorithmic systems and another in scientific institutions. They will demonstrate the ways in which trustworthiness by design can be implemented and what would be the advantages for our trusting relations that are the precious glue of our societies. Both case studies have been chosen in order to capture trustworthiness of the processes within the nexus of science, society, and policy in the context of the 4th industrial revolution.

In the first chapter, we dive all in by directly applying the trustworthy by design model to algorithms. This approach will enable us to directly test the model in the current rapidly changing environment of digital transformations that are taking ever tighter hold over human reality. We hold now is the right time, in this initial phase, to join together the development of digital technology and artificial intelligence with ethical considerations and impact they will have on society.

In the context of placing trust in institutions or digital platforms, we first test the optimism of trusting digital platforms by inquiring about the trustworthiness of algorithmic systems that run in the background. What interests us, in particular, are decision-making algorithms that exhibit big influence on decisions about our lives from health care, credit scores to legal systems. Can algorithms be trustworthy, can they make fair decisions? In the distributed systems of artificial and human agents who should be held responsible when something goes wrong? How to regulate algorithms if some of them are black box even to their designers? By mapping the most impactful decision-making algorithms we focus on the indicators that are influencing loss of public trust. That brings us closer to examining the impact of value-laden algorithms and accountability gaps

that are in particular present in machine learning algorithms and proposing the ways of ensuring algorithmic systems to be trustworthy by design.

The second chapter draws back to the origins of the concepts of trust and trustworthiness. We use critical analysis to unpack and closely follow the metamorphoses of trust from trusting relations between individuals, to forms of trust in institutions and transformations of trust in digital technologies. The aim of our critical analysis is to tackle the questions of trusting relations all the way to the origins of the concepts of trust and trustworthiness.

This pluridimensional map of concepts of trust and their unique context related characteristics should be a first step in tackling the questions of placing our trust in others, in institutions, digital platforms or algorithms and clarifying whether our trust is emotional or we can place it rationally. Based on the analysis of the state-of-the-art research of the concepts of trust and trustworthiness we propose three interrelated hypothesis. First, we argue that the relevance of the strong thin mode of trust has failed to be acknowledged but might be very beneficial in addressing the transformation of different modes of trust. Second, based on the elaboration of the unsustainability of the knowledge deficit hypothesis, we argue that both affective, emotional as well as cognitive aspect of trust play a significant role. Finally, based on the critical analysis of the different concepts of trust, instead of abandoning trust in institutions, we argue for a model that would better address identified shortcomings in the model of institutions trustworthy by design.

In the third chapter, based on the critical analysis of concepts of trust we formulate a model of trustworthiness by design that should be applicable both in institutions and in the context of digital technologies. The conceptual framework is based on the work of Onora O'Neill (2002a, 2002b, 2013, 2014) suggesting that in order to trust well we should place our trust in the trustworthy. We part from O'Neill in the view on the particular interrelation between emotional and cognitive aspects of trust and we broaden the understanding of trustworthiness so that it accommodates more inclusive and two-sided trust that places publics on the equal grounds. The proposed model of trustworthy institutions, therefore, rests on a three-part relation in which B is trusted by A for specific thing X, where we hold B to be an institution holding necessary properties of the institutional design. Here, institutions or systems are designed in order to accommodate responsibility and responsiveness next to usual properties of competence, honesty, and reliability. It reflects two-sidedness of trust between systems and publics by remodeling

power relations. Based on the critical analysis, we propose a revised form of mechanisms of openness and accountability. They should enable trustworthy institutions not only to signal its trustworthiness to the trustor but also to be responsive in a way to include uptake of trustees' potential concerns and values regarding the specific thing in question, in the communication *with* and not only *to* public.

Last two chapters test the devised model in the context of the scientific institutions. Trustworthy by design model accommodates participatory practices that in form of responsiveness and responsibility create space for achieving public trust by, at the same time, also trusting the publics. It accommodates not only signaling to trustor their trustworthiness but also includes responsiveness as an integral part, in a form of uptake of concerns and values that A holds regarding specific thing X. In the context of scientific institutions, this proposed model consequently implies participatory practices. In words of Peter Gluckman, that imagining of new modes of participation means that "building trust in science must start with building science capital … through deeper public involvement in the institution of science itself: framing questions, setting agendas, reviewing results or in other words better understanding of the concepts of co-design, co-production and extended peer review (Gluckman, 2017, p. 5).

Our main objective in chapter four is to start by mapping indicators that influence public mistrust in science. In order to do that and to tackle the question of trustworthiness of scientific institution in the nexus between science, society, and policy we place the discussion on trust in science within the broader context of science and values. We analyze the current debate on the value-free ideal of science. On one side proponents of the value-free Ideal in science (McMullin, 1982; Lacey, 1999; Mitchell, 2004) argue for excluding values from the scientific process. On the other side, within value-laden thesis the critics argue that there are several places for the positive role of non-epistemic values in science. One is based on the underdetermination argument (Harding, 1991; Longino, 1990, 2002) that science is underdetermined by evidence creating the gap for values to enter. Second is inductive risk argument (Rudner, 1953; Churchman, 1948; Douglas, 2000, 2009) where scientists confronted with inductive risk have to make value judgments on the sufficient evidence depending on the possible consequences of false positives and false negatives. Third place for values in science is articulated (Kourany, 2010) within the ideal of socially responsible science that introduces social as well as epistemic standards.

On this background emerge two approaches of addressing the question of trust in the context of scientific institutions that we analyze in turn, trust in individual experts and institutional skepticism approach. The closer analysis of those two approaches together with their application in the tobacco industry case study enables mapping the indicators that are influencing changes in trust in a scientific context. Based on the analysis we argue that two main indicators are influencing changes in trust. First is related to framing the problem and research agenda as well as neglecting diverse voices, often based on confirmation and macro-biases. The second indicator is closely related to the uncertain nature of science that could be used to delay regulations and have harmful consequences for the publics.

In the fifth chapter, we follow up on the identified challenges that indicate changes in public trust in science, arguing that trustworthiness of scientific institutions would offer a better prospect of addressing those issues of public mistrust in science. Similarly to the introductory chapter where we implement the trustworthy by design model to algorithmic systems, in this final chapter we test the same model in the context of scientific institutions. In contrast to institutional distrust model, we suggest a systemic redesign of scientific institutions as to accommodate principles of responsibility and responsiveness next to the properties of competence, honesty, and reliability within the model of trustworthy scientific institutions.

In order to encode those constitutive elements of responsibility and responsiveness, we draw back to two arguments that introduce positive roles of values in science, underdetermination argument, and inductive risk argument. We place them then in the broader context in which scientific institutions operate in the nexus between science, society, and policy, emphasizing the role of responsiveness and responsibility in relation to publics and democratic accountability when it comes to policy and policy advice. We then argue for extra-scientific critical contributions (Wylie, 2014) that directly relates to the publics and is an extension of the transformative criticism introduced by Longino (1990, 2002) originating in the underdetermination argument. Second, based on the inductive risk argument (Douglas, 2000, 2009), we develop a case for democratic accountability of the science policy advice, arguing that it could be achieved in an institutional setting by ensuring openness about value judgments and public participation.

CHAPTER 1: CAN ALGORITHMS BE TRUSTWORTHY?

When reflecting on the inadequacies of the trust in institutions, a suggestion to replace it with trust in digital technologies might be seen as a new prospect to achieve trustworthiness. And although it might be argued that trust in institutions should be replaced with trust in digital platforms, we suggest testing this claim by inquiring about the trustworthiness of the algorithmic system that runs in the background of the digital systems. To address the gap in the literature on the ethics of algorithms that do not pay sufficient attention to the aspect of trustworthy algorithmic systems we will apply our model of trustworthiness to algorithmic systems.

Shortages of trust in institutions are observed more closely in surveys and across the research fields. According to recent surveys (Edelman, 2017, 2018; European Commission, 2017), people are losing trust in institutions such as governments, businesses, media, and NGOs. The outlook on the modes of trust in institutions has been given attention in the philosophical literature (Govier, 1997; Hardin, 2002; Potter, 2002; Townley and Garfield, 2013), examining it through the lenses of power relations (Hardin, 2002; Cook, Hardin and Levi, 2005; Hardin, 2006) and concluding that trust in institutions is not necessary for cooperation (Cook et al., 2005). However, these analyses fail to acknowledge different modes of trust that take different forms in interpersonal relations than when implemented in an institutional setting. Those different modes of trust are also dependent upon how much we risk by trusting and if the stakes of our commitment are high or low.

According to the analysis of the changing trust modes in institutions introduced by Hosking (2014), a new mode of strong thin trust in institutions is becoming ever more predominant. This mode is strong because the decisions made in those institutions have a high impact on our lives, such as our financial stability or health. In spite of the high stakes our understanding of trusting relations is very low due to unavailable face-to-face contact in complex organizational settings that makes this mode thin. In combination, then strong thin mode of trust refers to a mode where our values are high but assurance of placing trust well in hindered. Hosking suggested that this strong thin mode of trust has been expanding for the last half-century as the radius where we place trust is ever bigger while at the same time more complex organizations and institutions hinder our ability to judge where to place our trust.

It is yet to be examined if this strong thin mode of trust in institutions causes or happens in parallel with new trusting relations driven by new digital technologies and whether trust in digital technologies can be a good substitute for trust in institutions. In this regard important research has been done on the topic of trust related to technology and artificial agents (Taddeo, 2010; Castelfranchi and Falcone, 2010; Buechner and Tavani, 2011) and recent related research on trust in algorithms (Cohen et al., 2014; Rubel and Jones, 2016; Shackelford and Raymond, 2014).

More specifically, when it comes to trust in a digital economy driven by new digital technologies Rachel Botsman (2017) argues that trust is currently changing its form. We now no longer place our trust upwards in institutions, but rather through networks or platforms within sharing economies such as Airbnb or Uber. This new distributed trust through technology at the same time changes dramatically the amount of trust we place in individuals and strangers. With new technological developments through blockchain technology where middleman is no longer required in the process, it might only gain more impetus in future co-creating practices.

Although these new developments in distributed trust enabled through digital technologies might be inspiring, here it is again worth remembering how fragile trust is. Hosking (2014) draws an analogy between trust and a coconut tree. He compares the growth of trust to a slow growth of coconut tree but also at the same time trust can be as easily destroyed as a coconut tree and once the trust has been broken it is not as easily repaired.

Taking into account the fragile nature of trust, we would want to consider both beneficial sides of digital technologies as well as potential challenges to trustworthiness. The development of new distributed trust through digital technologies at first also seems to fit more nicely with the requirements of placing our trust wisely which might make us more successful in trusting the trustworthy. Based on the O'Neill's (2002a, 2002b, 2013) account on trustworthiness, we can place our trust well only if we trust the trustworthy. In order to do that "placing trust in another for some particular purpose we typically need to make a judgement of the other person's competence and honesty, as well as their reliability, in the relevant matter" (O'Neill, 2013, p. 238). Judging the competences, honesty and reliability seem to only be enhanced within the platforms where we could consult ratings and reviews.

We could also more easily find the right fit and required expertise for a specific purpose that we are interested as envisaged in the three-part formula of trust where A trusts B for the

specific purpose X. It makes it easier to judge specific expertise that we require be it a babysitter, restaurant services or a dentist. Based on ratings and reviews we do not have to trust specific person generally for all purposes, but only for that specific purpose.

Together with benefits, digital technologies enabled by algorithms do not escape potential challenges. Sheer optimism, therefore, becomes tempered when we look closer into the processes distributed through digital technologies enabled by algorithms as the complete neutrality and objectivity quickly become questionable. Personalization and filtering algorithms (Hildebrandt, 2008; Newell and Marabelli, 2015; Taddeo and Floridi, 2015) for example are beneficial in terms of filtering only relevant information for a specific user. The downside however of making decisions on what is important for users is at the same time reducing diversity (Barnet, 2009; Pariser, 2011) and having a direct consequence on the user autonomy (Newell and Marabelli, 2015).

Algorithms raise ethical together with epistemic concerns as can be illustrated in practical examples of Facebook personalization algorithm or profiling algorithms (Mittelstadt, Allo, Taddeo, Wachter and Floridi, 2016). The Facebook personalization algorithm that prioritizes what content will appear on the top of the page is based on the parameters that can be interchangeable, depending on what is decided to be more relevant, such as date of the content, type of the relationship with the person who published the content, adds, media, etc. Well-known echo chamber effect can be a direct result of reducing the diversity of information we are receiving.

Another often-used algorithm is the profiling algorithm that bases predicting behavior of individuals based on connections he or she has with other members of a specific group and can thus lead to discrimination. Those concerns regarding our autonomy and discrimination are some of the ethical concerns that go hand in hand with further epistemic concerns raised by outcomes of some algorithms (Mittelstadt et al., 2016). These concerns become more tangible when algorithms are used for the important decisions that will influence our life in deciding on being granted credit, having job opportunities, going through a criminal system, etc.

Those among other policy concerns have been the topic of extensive recent interest (Pasquale, 2016; Citron and Pasquale 2014; Zarsky, 2013; Crawford and Schultz, 2014). Such ethical and legal concerns lead to questioning the trustworthiness of algorithms in making decisions about important aspects of our lives, our health, jobs or financial prospects. Could

algorithms be trustworthy? Trusting algorithms and holding them responsible seems like an effective way of riding humans from responsibility, deferring to algorithms and trusting their automated processes (Zarsky, 2016, p. 121). Can we trust algorithms to make the right decisions for us? And if that process fails who should be responsible?

Addressing those questions in the first part we will start with examining the impact and implication that decision-making algorithms have for society. Mapping the most impactful decision-making algorithms and addressing related ethical challenges would give us an indication of where changes in trust may lay. As algorithmic design cannot be separated from the societal context the first challenge to address is value-leadenness of algorithms. Second, we address the question of how to assign responsibility and accountability. Should it be addressed to designers of algorithms, their users or algorithms themselves? Although the answer might seem straightforward, it tends to complicate when it comes to learning algorithms that pose evident autonomy and can be a "black box". That leads us to the question whether can algorithms be trusted, can they be responsible and accountable or is that capacity inherent only to human agency?

In the second part, we place the discussion of algorithmic systems in the broader context of digital transformations. Should those new and fast developing transformations in the digital sphere trigger new forms of regulations? And how does that relate to regulation of algorithms? Is their regulation necessary for maintaining our trust or it unnecessary stifles innovation? In this context, we address the challenges of transparency, in terms of accessibility that is not granted due to secrecy reasons and in terms of comprehensibility related to the black box aspect of machine learning algorithms. In the final instance, we examine three proposals for regulating algorithms questioning if they are to provide sufficient ground for the trustworthiness in the context of algorithms.

In the final part, we examine the question of trustworthiness of algorithms and propose our model of trustworthiness by design illustrated in blockchain technology. It should be better suited to address this complex interrelation between human and artificial agents in algorithmic systems and organizations that are designing and implementing decision-making algorithms.

1.1 Decision-making algorithms, values and responsibility

Decision-making algorithms are embedded in the society directly influencing our systems of law, health, finance or policing. The benefits they are bringing in terms of effectiveness, enhanced capabilities and potential for escaping human biases are coupled with implications for fairness, discrimination, emphasizing biases and questioning responsibility. Implications of decision-making algorithms have a direct societal impact. Therefore, we claim that critical role for the prospect of algorithmic systems trustworthy by design should be considered within the prospect of responsibility, responsiveness, and diversity implemented through participatory, and inclusive practices.

These participatory and inclusive practices should be at place from the very early phase of algorithmic design and throughout the process of implementation. Our intention in this section is to map the most impactful decision-making algorithms and to address ethical challenges they raise. That will serve as a starting point in identifying indicators that influence change in trust in the context of algorithmic systems. Based on that analysis we can then devise a model of trustworthiness by design with the aim of addressing identified emerging ethical concerns.

In addressing decision-making algorithms we first have to clarify what we consider under algorithms. We will adopt the definition of algorithms that are not only mathematical constructs but also implemented in "computer programs, software and information systems (…) as mathematical constructs, implementations (technologies, programs) and configurations (applications) (Mittelstadt et al., 2016, p. 2). This extended view about algorithms corresponds closer with the public discourse of understanding algorithms instead of limiting their understanding only within the mathematical construct.

What specifically interest us, however, are decision-making algorithms that are used in various situations and differ largely in complexity. They range from already mentioned personalization algorithms (Hildebrandt, 2008; Newell and Marabelli, 2015; Taddeo and Floridi, 2015) and profiling algorithms (Hildebrandt, 2008). Machine learning algorithms, that are not solving problems according to predefined programming but are learning how to solve problems and therefore are unpredictable (Tutt, 2016; Burrell, 2016), negotiation algorithms helping navigate pro and cons in negotiation processes (Raymond, 2015) or clinical decisions algorithms in computer-based diagnostic systems (Mazoue, 1990).

The benefits and challenges those algorithms bring in the process of decision-making are abundant but also bring various challenges. Among them, we focus on two main challenges. First is the value-leadenness of algorithms, the overarching concept under which we include biases that influence discrimination, a question of fairness, diversity, and inclusiveness. Second, and the closely related issue is about responsibility related to governance challenges, regulation, accountability, transparency, public engagement, and finally trustworthiness by design.

1.1.1 Value-laden algorithms

Algorithms improve decision-making processes, improve efficiency, produce new knowledge and often can be seen as being free of human biases. They are already embedded in so many aspects of our lives, from selecting playlists, news feeds, to health, financial, legal decisions. Designing algorithms in the context of computer systems, however, is also value-leaden (Brey and Soraker, 2009; Nissenbaum, 1998; Flanagan, Howe and Nissenbaum, 2008). Friedman and Nissenbaum indicate several points where values can appear in computer systems (Friedman and Nissenbaum, 1996). The first is at the point of design when the biases from the designer or his institution enter into the very design. These could also be called "pre-existing biases". The second type of biases is related to particular restraint imposed by the technology used. And the third is in the process of its implementation that might happen in the context that was not envisaged.

Recently, this value-laden aspect of algorithms has been looked at more closely (O'Neil, 2016; Barocas and Selbst, 2016) provoking the discussion about the values behind as algorithms proved not to be value neutral. Examples of biases within machine learning processes recently got more prominence as they are increasingly being implemented for decision in various aspects of our lives. Therefore, they can have tangible implications in ethical and legal sense for different social groups based on racial, gender or other biases.

Racial biases have been addressed in research as bias in search algorithms that exposes racism discriminating people of color (Noble, 2018), racially biased data in predictive policing (Lum and Isaac, 2016; Ferguson, 2017), racial differences that influence participation in clinical trials (Kurt et al., 2016), misdiagnoses based on the studies that exclude black (Lacy et al., 2017),

etc. One of the prominent examples of racial biases was examined in ProPublica (Angwin et al., 2016) on the decision making process on recidivism of prisoners that has been discovered to largely discriminate black.

The ProPublica study was conducted on 7000 risk scores on risk assessment of future crimes that were handed to judges during the sentencing process across the US in order to test the accuracy of the algorithms used. It showed racial disparities as recidivism (the prediction of repeating the crime) was falsely assigned to black (44.9%) almost twice the rate of white (23.5%) defenders while white defenders (47.7%) were more often falsely freed of the risk of repeating a crime than black defenders (28.0%).

The risk assessment tool that was used, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) next to risks also assessed other aspects related to criminology such as "criminal personality", "social isolation", "substance abuse" or "residence/stability". The company's software used for the risk assessment correctly predicts recidivism 61 percent of the time. They, however, reject the ProPublica analysis and for proprietary reasons keep secret the code used for calculating the scores thus making it harder to challenge the decision.

Examples of gender biases, on the other hand, have been found in ads targeting male more than female for high paid jobs (Datta et al., 2015; Campolo et al., 2017) or for careers in STEM although they are originally intended to be gender neutral (Lambrecht and Tucker, 2016). Furthermore, machine-learning algorithms use word embedding that perceives words as vectors and understands them based on their relation with other words. It is often used for ranking in Web search (Nalisnick et al., 2016) or for resume analysis (Hansen et al., 2015), etc. Closer analysis (Bolukbasi et al., 2016) shows how word embedding can also exhibit gender stereotype and gender biases as it has been found that *man,* for example, appears more often next to *computer programmer* and *woman* next to *homemaker.*

Far from being morally neutral the very design of algorithms is situated in the social context and is inevitably influenced by values and biases. Addressing the social aspect of machine learning processes behind the digital technology points to the question of inclusive algorithms and their attainability. Furthermore, biases amplified by machine learning algorithms could have an even greater societal impact. Promulgating unfair outcomes in terms of decisions about our health, credit scores, jail sentences etc. becomes unsustainable as it gets bigger in scale.

In the final instance, it can result in the mistrust on behalf of publics and citizens that are being subjected to unfair systems.

In regards to the value-leadenness of algorithmic design Cathy O'Neil (2016) further elaborates on the relation between models and values. She defines models in simplest terms as an abstract representation of some process used to predict outcomes. Creating models includes value judgments throughout the process of modeling from its purpose, a decision on the input to the ways we define favorable outcome. She further illustrates how creating the model resembles an analogous process of preparing food. In both you have to decide what goes in, what data or food is decided to be important enough to be included, questions we ask and the definition of success. The illustrative example is also Google maps when asked for the direction they do not include buildings and trees because it is decided that is not important for the purpose of the question. All those particular decisions reflect priorities that are built into the model. Therefore, models cannot be considered impartial and value-free application of mathematics.

O'Neil (2016) further reveals how biases are built into the software systems emphasizing their implications in our democratic society:

> "The math-powered applications powering the data economy were based on choices made by fallible human beings. Some of this choices were no doubt made with the best intentions. Nevertheless, many of these models encoded human prejudice, misunderstanding, and bias into the software systems that increasingly managed our lives. Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, were beyond dispute or appeal. And they tended to punish the poor and the oppressed in our society, while making the rich richer." (O'Neil, 2016, p. 3).

This model of possible destructive impacts on the society she terms "weapons of math destruction".

Value-laden algorithmic design that builds desired outcomes into the algorithms (Kraemer et al., 2011) could cause discrimination and unfairness. One of the most often examples of discrimination is against the marginalized population in profiling algorithms in particular. In examining discriminatory practices Barocas and Selbst (Barocas and Selbst, 2016) focus on the

case of American Anti-discrimination law on the prohibition of discrimination of employment warning on the possible consequences, because

> "Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court." (Barocas and Selbst, 2016, p. 671)

In their essay they are not calling against data mining but are warning that there are cases where precaution is needed because of the harmful effects it might bring;

> "Ideally, institutions can find ways to use data mining to generate new knowledge and improve decision making that serves the interests of both decision makers and protected classes. But where data mining is adopted and applied without care, it poses serious risks of reproducing many of the same troubling dynamics that have allowed discrimination to persist in society, even in the absence of conscious prejudice" (Barocas and Selbst, 2016, p. 732).

Their call for precaution is rightly addressed to institutions, as they together with other private and public actors are the best enabled to anticipate possible risks of mistrust. In that quest responsibility in preventing discriminatory and unfair practices in an open and accountable way also contributes to their trustworthiness.

1.1.2 Who is responsible?

In tackling the question of responsibility for the potential problems and harmful societal impacts of algorithms determining who should be responsible and accountable may not be as straightforward as it might seem. Assigning responsibility becomes even more challenging since it is often distributed between several actors or it has often be applied to machine learning algorithms.

According to our previous overview of values entering in the computer system at the point where algorithms are designed where it is decided what data goes in, what is the purpose and what is the desired outcome. In that case, the question of responsibility should be directed to the

actors who are designing algorithms. The designer behind the algorithm should be responsible for designing technology for a specific purpose, analogues to the responsibility of the software designer who can explain and indicate potential risks (Floridi at al., 2015). At the point of implementation, users should be blamed and held responsible. At least this is the first way in which we could assign responsibility to designers of algorithms, as "software designers are morally responsible for the algorithms they design" (Kraemer et al., 2011, p. 251). The third place where bias can happen regarding the constraints of technology should also be considered and assign responsibility accordingly.

However, the complex process of designing, regulating and using algorithms for decisions that are influencing our lives on a daily basis from health, jobs, to schools can no longer be seen through frameworks focused only on individual responsibility. The process is becoming more complex, involving various actors from software developers, designers, users, regulators, etc. Therefore, instead of individual responsibility new ethical theories on distributed agency (Floridi, 2013; Floridi and Taddeo, 2016) are better suited to hold "all agents of a distributed system, such as a company, responsible. This is key when considering the case of AI, because it distributes moral responsibility among designers, regulators, and users. In doing so, the model plays a central role in preventing evil and fostering good, because it nudges all involved agents to adopt responsible behaviors" (Taddeo and Floridi, 2018, p. 751).

It becomes even more obvious when it comes to learning algorithms when the amount of prediction on behalf of the designer becomes questionable because algorithms possess certain autonomy. Machine learning algorithms, therefore, pose new challenges (Burrell, 2016; Zarsky, 2016). It can lead to an accountability gap (Cardona, 2008; Mittelstadt et al., 2016) as designers no longer have an exact overview of the potential unintended consequences or harm that algorithm behavior might produce. It is specifically salient when it comes to learning algorithms such as for example genetic algorithms that are programming themselves. Determining flaws in learning algorithms poses a unique challenge because of the uncertain nature of resulting learning processes and their complexity.

Machine learning algorithms as a basic part of their learning process to some degree autonomously define decision-making rules. That makes it harder to predict how they will deal with new inputs or to explain it afterwards due to the complexity of the process. Consequently, it also complicates the assigning the responsibility. If a problem occurs who should be held

responsible, the designer of a semi-autonomous algorithm or the algorithm as a moral agent? Mittelstadt et al. (2016) conclude that it further complicates the problem of implementation of ethical principles in automation and application of the distributed responsibility across the network of human and algorithmic actors.

The second challenge that machine learning algorithms pose in assigning responsibility and accountability is related to their comprehensibility (Mittelstadt et al., 2016). They can be a "black box" even to its designers and pose the challenge of correcting them in the real-time either due to informational advantage or the processing speed of algorithms. Moreover, large teams often develop them over a period of time which makes almost impossible to trace biases. The rationale of algorithms whose oversight might be hindered as they might be incomprehensible to humans makes them untrustworthy and challenges the quest for assigning responsibility and accountability.

Burrell (2016) places the features of opacity in learning algorithms in the context of their real-world societal implication where they directly influence decisions in different areas of our life

> "as a problem for socially consequential mechanisms of classification and ranking, such as spam filters, credit card fraud detection, search engines, news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring. … These mechanisms of classification all frequently rely on computational algorithms, and lately on machine learning algorithms to do this work." (Burrell, 2016, p. 1)

Apart from other possible reasons for opacity due to commercial competitiveness, secrecy or technical illiteracy, Burrell (2016) focuses on the potential implications of machine learning algorithms where "rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs" (p. 1). They, therefore, indicate a specific opacity challenge

> "at a more fundamental level. When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension. Machine optimizations based on training data do not naturally accord with human semantic explanations. The examples of handwriting recognition and spam filtering helped to illustrate how the workings of machine learning algorithms can escape

> full understanding and interpretation by humans, even for those with specialized training, even for computer scientists" (Burrell, 2016, p. 17).

As a consequence, decisions that might have an effect on the people can be hardly comprehensible and understandable. Consequently, it calls into question the legitimacy of placing trust in them and also concerns that such decisions could hardly be challenged. The accountability problem here becomes more profound requiring further specifications in regards to the question of trust. Who is the subject of trust, for what purpose, in which specific context?

If algorithms are to be held accountable, do they hold artificial moral agency related to the human moral agency? The autonomy of learning algorithms actually can be understood as a part of a moral agency, but the one that is different from the human moral agency. Floridi and Sanders include autonomy to be a part of the moral agenthood together with interactivity and adaptability when they set guidelines as "interactivity (response to stimulus by change of state), autonomy (ability to change state without stimulus) and adaptability (ability to change the 'transition rules' by which state is changed) at a given LoA" (Floridi and Sanders, 2004, p. 1).

However, the role of a moral agent in this case in terms of accountability is strictly to be distinguished from responsibility that is inherent only to human agents due to capabilities of intentionality. This approach initiates the discussion about the ways of enabling oversight in a more complex entanglement of human and artificial agents. It abandons the one-sided approach of either hiding behind the algorithms that should be blamed when something goes wrong or directing the blame to bad design (Anderson and Anderson, 2014).

Although algorithms could be held accountable, they could not be trusted. Trust can be directed only to the human agency that has an oversight over the process as the only agent able of intentionality and consequentially moral responsibility. Therefore, trust in algorithms (Zarsky, 2016, p. 121) that could be attributed no intentionality would be misguided as well as the question about trustworthy algorithms that is also displaced. This complex interconnectedness of accountability of algorithmic agents and responsibility of human moral agent capable of intentionality has to be taken into account when conceptualizing the prospect of oversight and trustworthiness of algorithmic system.

1.2 Should algorithms be regulated? Can they be transparent?

Our discussion of trustworthiness of algorithmic system and analysis of the prospect of assigning responsibility and accountability for the automated decision-making processes that can have a substantial impact on people's lives is part of a much broader rapidly changing environment that some call the fourth industrial revolution (Schwab, 2017). In this rapidly changing environment of digital transformations, emerging technologies, such as machine learning, AI, the blockchain, big data, the Internet of Things, etc., will have an immense impact on changes in society both positive and negative. From a positive perspective, new transformations might enrich the quality of our lives, healthy living and aging. At the same time, there are early warnings on the risks and negative aspects that should be anticipated such as job losses due to automated processes, surveillance practice, serious impact on inequalities, etc.

The broader question then leads to how to enable protecting the citizens from potential harms. Should new regulatory mechanisms be devised similar to the regulations that ensure the safety of the food that we eat, our health against exposures of chemicals or safety regulations in the environment? Or should the innovation not be stifled and industries that are in the forefront of the new revolution should be given free hands? Have we learned from the previous revolutions the effects that innovations can have on the public's could also cause serious harms that could take years to be mended? So, what are the prospects in the new revolution for the citizens and the publics?

Our discussion on trustworthy algorithmic systems by design is just one of the segments of this broader picture of current digital transformation landscape. Should algorithms be regulated so that we can trust them? If yes, how exactly should the regulations take place? The new technological sphere brings with it a new *modus operandi* that is producing new models and innovations much faster and at much greater scale. New developments, therefore, require more adaptable and responsive regulations.

But when it comes to the specific question of regulating algorithms the challenges of access to information about them as well as the understanding of their workings poses a challenge to regulatory options. Although it might be questioned if the transparency can solve the ethical issues raised by algorithms as “transparency is no guarantee of legitimacy” (Crawford, 2016, p.

7). Furthermore, opening up and disclosing data and information might be rightly questioned as limited and unsatisfactory in this regard.

But even if we want to enable this first step of transparency for the regulatory process that would be applied to algorithms, there are two aspects of transparency that might be hard to achieve in the context of algorithms. Formulation of the way in which transparency is usually used and understood could indicate where the challenge lies. The transparency is often defined

> "from the perspective of those who gain access to information (e.g. the public, employees or regulatory bodies), transparency depends on factors such as the availability of information, the conditions of its accessibility and how the information, which has been made transparent, may pragmatically or epistemically support the user's decision-making process. Information providers (e.g. companies, organisations or public institutions) shape such factors by choosing which information could or should be disclosed, also according to current legislation, and by deciding in which form information might be most suitably made available. Such choices and decisions depend on evaluating business, legal and ethical constraints and implications." (Turilli and Floridi, 2009, p. 106).

Transparent information that is available and accessible, nevertheless, might be unintelligible, without understanding how and by whom it was produced, and not communicated in a comprehensible way. According to Turilli and Floridi transparency understood as only putting the information into open can be perceived as too limited

> "The common understanding of information transparency as the process of disclosing a set of data has been challenged as too limited, in favour of a more inclusive definition that takes into account also the ethical principles factually endorsed in producing information. On the basis of such a definition, we have argued that disclosing not only information but also details of how such information has been produced enables those ethical principles that either depend on information or regulate it (Turilli and Floridi, 2009, p. 111).

If we understand transparency in terms of its two main features "accessibility and comprehensibility of information" (Mittelstadt et al., 2016) the prospect of transparency when it comes to algorithms confronts two main challenges. The first one is related to the prospect of accessibility that might not be available due to secrecy in order to protect commercial competitiveness in the context of industry, or privacy and national security. The second challenge to comprehensibility is related to the black box problem specifically when it comes to the

machine learning algorithms where the rational of devising the solution and the decision can often be incomprehensible even to the designer of the algorithm. We will focus on each of those two aspects in turn in order to envisage a possible prospect for the algorithmic trustworthiness by design.

First, accessibility is often intentionally not granted and secrecy is being kept for several reasons (Mittelstadt et al., 2016), it might be either for national security reasons or privacy of data subjects or organizations autonomy. In the industrial context it is justified due to competitive advantage as disclosing the algorithms structure would make it possible to game the system and for example manipulate search results.

> "Coding often happens in private settings, such as within companies or state agencies, and it can be difficult to negotiate access to coding teams to observe them work, interview programmers or analyse the source code they produce. This is unsurprising since it is often a company's algorithms that provide it with a competitive advantage and they are reluctant to expose their intellectual property even with non-disclosure agreements in place. They also want to limit the ability of users to game the algorithm to unfairly gain a competitive edge." (Kitchin, 2017, p. 20)

The secrecy, on the other hand, hinders people that were subjected to the decisions based on the secret algorithms to understand the very process that was responsible for the results.

At this point of relation between transparency and how it relates to the public interest, O'Neil (2016) pushes the discussion even further identifying secrecy as one of the aspects of WMD (Weapons of Math Destruction) together with scale in terms of scalability of the models and their application to large number of people and damage that they can cause. She frames opacity in industries as "secret sauce" that treat algorithms as property rights creating "black boxes" and thus posing a question about public interest.

This condition of opacity Pasquale (2016) observes further in terms of possibly deliberately stimulated within the frame of deliberate ignorance. This opacity as he suggests might offer a veil against regulation; "But what if the "knowledge problem" is not an intrinsic aspect of the market, but rather is deliberately encouraged by certain businesses? What if financiers keep their doings opaque on purpose, precisely to avoid or to confound regulation? That would imply something very different about the merits of deregulation" (Pasquale, 2016, p. 2).

A very illustrative example of such a secret sauce and black box algorithms is in the recidivism bias when states estimate future possible crimes of the released prisoners based on the algorithms. As algorithms are in the software developed by the private companies, their internal workings often will not be available to judges who will give the sentence based on that information. ProPublica recently investigated in detail one such a case elaborating on the not completely open access to the code, although some states did evaluate separately accuracy of statistical analysis they've used (Angwin et al., 2016).

One of the inspiring practices has been recently implemented in the New York City as to address the potential societal consequences of the algorithms that might influence unfairness and inequality. As declared on the official website ("Mayor de Blasio Announces", 2018) the NYC has formed a task force to develop procedures for assessing algorithms that city uses in decision-making process across departments be it on social services, schooling or criminal justice. The automated decisions will be assessed for fairness, equity and accountability by the diverse task force that brings together members from the city departments together with academics, legal and data experts, nonprofits and think tanks. Decisions made by algorithms should, therefore, be accessed by these task force in order to ensure that they will take into account their social impact and effect on different groups.

Comprehensibility is the second aspect of transparency that we have already referred to as a "black box" problem of machine learning algorithms. Due to their semi-autonomous nature, they can be opaque even to their designers (Mittelstadt et al., 2016). The EU General Data Protection Regulation in force from 25 May 2018 is a recent example of the challenges to a transparency of decision-making algorithms. Data controllers now have an obligation to estimate the risks and communicate them in a plain language and explain the logic behind the decisions to data subjects for example in profiling activities (Art. 12, 13,14, 35). This aspect of explaining the automated decisions made by algorithms is challenged again by the comprehensibility requirement. It might be difficult to satisfy "explainability" criteria specifically for data industries due to "the connectivity and dependencies of algorithms and datasets in complex information systems, and the tendency of errors and biases in data and models to be hidden over time" (Mittelstadt et al., 2016, p. 14). Following on the diverse line of arguments on the attainability of explaining algorithms in the context of algorithmic decision-making processes even if we take

them not to be sufficiently explainable, there is one more important question that we still have to address.

Even if it would be accepted that the competences and exact working of the algorithms are not sufficiently explainable and thus lack comprehensibility as an integral part of transparency, a more important question still comes back to the human intentionality behind the process and ensuring its reliability in the line with public expectations. This responsibly must entail oversight of the process behind the design and application of algorithms, the option of reversing or modifying the process and ensuring that the outcomes are in line with the public interests. As it is also emphasized in the proposed mapping of the ethical aspects of algorithms that

> "a perfectly auditable algorithmic decision, or one that is based on conclusive, scrutable and well-founded evidence, can nevertheless cause unfair and transformative effects, without obvious ways to trace blame among the network of contributing actors. Better methods to produce evidence for some actions need not rule out all forms of discrimination for example, and can even be used to discriminate more efficiently. Indeed, one may even conceive of situations where less discerning algorithms may have fewer objectionable effects." (Mittelstadt et al., 2016, p 14-15).

Even in the case of not completely attainable transparency, ensuring that the outcomes are fair or could be reversible plays a crucial role. But aspects of responsibility usually do not come to the fore straight away in delegating power to algorithms and digital technology. However, when processes go wrong and outcomes are dangerous or highly disturbing, the question of responsibility becomes highly relevant since someone has to take responsibility.

1.2.1 Overcoming accessibility and comprehensibility challenges

There are, however, suggestions for regulating algorithms that try to overcome the accessibility and comprehensibility challenges. For the purpose of our discussion, we will categorize them in the three categories. First is the proposal for external regulating bodies or institutions, second is the internal regulating mechanism that would be enabled by algorithms that are explaining algorithms. And the final regulation option emphasizes the human aspect in the process and the relevance of experts that have domain-specific knowledge in the regulatory topic.

Solutions that would be apt for holding machine learning algorithms to account are often suggested as a "combination of regulations or audits (of the code itself and, more importantly, of the algorithms functioning), the use of alternatives that are more transparent (i.e. open source), education of the general public as well as the sensitization of those bestowed with the power to write such consequential code" (Burrell, 2016, p. 17).

First suggestion refers to regulation by an independent external regulator (Pasquale, 2016; Zarsky, 2016; Tutt, 2016). Overarching initiatives on the external regulation of algorithms propose independent body or institution as regulators of algorithms. Pasquale (2016) suggests one of the potential solutions to overcome the opacity of algorithms by auditing the code by the independent third party. That solution should both satisfy the quest for secrecy due to competitive advantages in the industry and regulate potential discrimination or another potential misuse. Along those lines, a more specific suggestion is to create centralized federal agency for regulating algorithms (Tutt, 2016) that would be better suited for this purpose than subject-specific agencies. The proposal of FDA for algorithms is based on the analogy of potential harms that algorithms could cause to the harms of contaminated food, unregulated drugs or cosmetics,

> "Algorithmic regulation will require federal uniformity, expert judgment, political independence, and pre-market review to prevent - without stifling innovation - the introduction of unacceptably dangerous algorithms into the market. This paper proposes that a new specialist regulatory agency should be created to regulate algorithmic safety. An FDA for algorithms." (Tutt, 2016, p. 83)

Contrary approach on internal initiatives is regulation by algorithms. It suggests developing algorithms with a purpose of regulating, tracking and explaining the algorithms, such as the proposal of different ways of interpretable machine learning (Vellido et al., 2012). Although the prospect of this solution of algorithms explaining and understanding algorithms, seems like an easy solution in handing over responsibility and blame to artificial agents, as we have argued, because intentionality is inherent only to humans, responsibility also can be assigned only to humans. What is unclear in this approach is whether the regulation within the industry itself that is creating algorithms to explain algorithms effectively achieves the prospect of regulations in line with the public's and citizen's needs and requirements. The newly introduced approach of the DARPA (Defense Advanced Research Projects Agency) on

explaining algorithmic decision-making processes might offer an insight into developing new ways of explainability that will advance oversight in using algorithms.

The third way of unpacking black box algorithm workings involves experts. Kitchin (2017) proposes this approach of giving the role to experts when he elaborates on reverse engineering. Instead of solely relying on the internal mechanisms of algorithms explaining themselves, understanding algorithms workings also "requires that the researcher is both an expert in the domain to which the algorithm refers and possesses sufficient skill and knowledge as a programmer that they can make sense of a 'Big Ball of Mud" (Kitchin, 2017, p. 23). This point of including experts of different backgrounds lead us to our next point that places the regulatory question much closer to the designing process of algorithms that is closely related to the question of the trustworthiness of algorithmic systems.

1.3 On trustworthy algorithms

If we want to answer the question whether we can trust algorithms and whether algorithms could be trustworthy we should examine it within the concept of trustworthiness (O'Neill, 2014) in the three-part relation where A trusts B for a specific thing X, we have to be able to judge competence, reliability, and honesty. Starting with data processors, provided transparency might enable our trust in data subjects (Cohen et al., 2014; Rubel and Jones, 2016; Shackelford and Raymond, 2014), although the practical attainability might be hindered through opacity (Mazoue, 1990). Trust can also exist between actors in the distributed system (Simon, 2010; Taddeo, 2010). But apart from data processors, could we place trust individually into algorithms?

Here we come to the second crucial point of O'Neill's (2013) concept of trustworthiness that is not attitudinal but centered on the intentionality. Intentions of algorithms cannot be discovered if they are taken as objective, automated systems instead of being designed in the process that cannot completely be value-free. The aspect of trustworthiness poses a further challenge in the context of machine learning algorithms that as agents have a specific amount of autonomy and subsequently could be held accountable. Although algorithms could be held accountable, they do not have a capacity for intentionality, which is inherent only to human agents. Therefore, the question of trustworthy algorithms is misplaced. Instead of searching for

trustworthy algorithms, trustworthiness should be related to complex interconnection between human and artificial agents. Since humans have intentionality and capacity to be held responsible and the ability for oversight of the artificial agents within the system that are to be held accountable.

Model of trustworthiness that could be applied to different agents, be it organizations, relations between humans, artificial agents or their combination should be based on judging the competence and reliability in the context of underlying intentions. However, next to only signaling (Jones, 2013) trustworthiness to the trustor, we suggest that the model also have to accommodate responsiveness in terms of uptake of trustees potential concerns and values that is best done through co-design processes.

Reframing the question of trustworthiness of algorithmic system should place it early in the process of algorithmic design. Within Crawford's (2016) proposal of agonistic algorithms, she refocuses the discussion from main functionality of algorithms to the broader perspective of the places where their design takes place.

> "How else might agonism be useful to us when thinking about algorithms? Rather than fetishizing the algorithm itself, theories of agonism allow us to widen the perspective to include the contested spaces where algorithms are designed. They are always made by and in relation to people: they are in flux and embedded in hybrid spaces. Thus, we can look to the companies and offices where algorithms are created as fruitful sites of research. These workplaces are themselves spaces of everyday conflict and dissent, where algorithmic design decisions are made after debate, disagreement, tests and failures. And we can look to the spaces where algorithms and people are interacting in quite public ways: for example, Reddit makes part of its algorithmic ranking process public. Users like having more awareness of the rules of the system, and some enjoy the possibility of gaming it, collectively or individually. While transparency is no guarantee of legitimacy, it is does mean that more opaque, autocratic systems (such as Facebook) generate more suspicion." (Crawford, 2016, p. 7)

It might further open the discussion on the spaces such as institutions and companies where algorithms are designed in the interplay between people and algorithms. This interaction spans to further spaces of algorithmic applications where interdependence between algorithms and people takes on a more impactful role as decisions that influence people's lives might be delegated to

algorithms. Due to that fact this perspective on the institutional places of their creation point in the right direction when contemplating on prospect the of trustworthiness.

What those co-design processes would look like in the context of algorithms may be clearer when illustrated through a practical example of civic engagement. One recent step in this direction is a new Gobo tool that is designed to empower and include citizens in co-designing a feed they want to see on social media, instead of being subject to black box workings of the algorithms behind the decisions of which post has a priority when showing on our feed.

This open source tool basically aggregates social media (the Twitter and available parts of Facebook) attaching then algorithms to them. But the main point is that it counteracts black box algorithms, they are open and the user can control filters and decide which type of post they want to give priority and they can also write filters for the tool itself - in that way to have co-creative control over the feed they see. The option "lot of perspectives" exposes to feeds from different sources that might not be our first choice and in that way counteracts echo chamber bubble. Gobo tool is created by the Centre for Civic Media at the MIT Media Lab and Comparative Media Studies at MIT.

That is one of the illustrative examples of the co-designing processes that contribute to the prospect of trustworthiness. However, its impact and implications should be scaled up to organizational agents that are designing and implementing algorithms in a way that their responsibility and accountability would be ensured. There are highly inspiring and promising steps in this direction (Campolo et al., 2017), however, further research in this field that would focus on participatory processes is needed.

1.3.1 Algorithms trustworthy by design

Neither of the discussions about two big tickets that we have analyzed, value-laden algorithms or the outlined proposals for the regulation of algorithms, do not pay sufficient attention to participatory processes in algorithms design or the relevance of inclusion and diversity of actors who will be influenced by the automated decisions. In order to emphasize the importance of that perspective that is not currently sufficiently addressed we suggest reversing

the focus from consequences of the algorithmic decision making to the earlier phase where design takes place by introducing algorithms trustworthy by design.

The results of the processes in detecting biases, for example, are mainly traced back only after they happen. The regulatory approaches of algorithms are mostly limited to providing explanations and interpretations of algorithms not taking into account participatory realm. However, this shift in the perspective was introduced by Sheila Jasanoff (2016) in the context of the ethics of inventions related to information technologies. She emphasizes the concept of prevention and possible alternative pathways that should be taken into account early on in the process of design in order to align it with the needs of the publics and not necessarily for the sheer attractiveness of technological advances or because we are able to produce it. In this way, it is possible to place the process of an invention within a more inclusive and democratic context. The application of this mode of thinking in the context of algorithms up to now has still not been sufficiently explored.

In order to better address participatory and inclusive processes from the early design phase, we suggest a new model that should be trustworthy by design. The model has two different features, first is related to encoding trustworthiness in the very design of algorithms and the second aspect is related to decisions on the purpose of using specific technology or algorithms. The term trustworthy by design integrates properties of the model of trustworthiness that we will develop in detail in chapter three. This model applied to algorithms enables developing technical functions so that they are an integral part of the autonomous decision-making process. It further implies reframing the technocratic division between ethical and social values on one side and science and technology on the other side (Smallman, 2018) by rather integrating them as an integral part of developing algorithms.

Prerequisite for algorithms trustworthy by design is based on recognition of the source of moral responsibility as inherently human, based on intentionality and originating in human values. Furthermore, it is recognizing the human framework within which algorithms are designed that enables oversight through mechanisms of accountability, openness, and audit. At the same time, it ensures responsible conduct and responsiveness as uptake of values of trustor in regards to specific application in question.

Together with integrating trustworthiness in the design of the algorithms, second equally important aspect is related to the decisions on the purpose for which machine-learning algorithms

will be used. One illustrative example is a recent study on risk assessment in criminal justice that suggests reframing the approach of using machine learning in this context. Their analysis shows how predictive risk assessment technologies actually

> "fuel harmful trends towards mass incarceration and growing inequality in the justice system. Predictive risk assessments offer little guidance on how to effectively intervene to lower risk. When predictive accuracy is the primary metric along which these technologies are evaluated, the system misses opportunities to explore a deeper set of questions surrounding the way its administrators can use data as part of a reflexive practice of testing hypotheses in the service of achieving near and long term goals" (Barabas et al., 2018, p. 7)

Therefore, they "argue that risk assessments should be conceived of as a diagnostic tool that can be used to understand the underlying social, economic and psychological drivers of crime" (Barabas et al., 2018, p. 7). Instead of perpetuating growing inequalities in justice system by applying predictive policing, machine learning algorithms could instead be used as tools for understanding what drives crimes in the first place.

There are already in place initiatives that greatly contribute to making algorithmic systems accountable. One of the suggestions is Algorithmic Impact Assessment AIA (Reisman et al., 2018) that would be conducted before governmental agencies employ automated decision mechanisms. The AIA would address potential risks for fairness and justice by engaging community, researchers and experts and thus enabling accountability of the process.

Although the proposal is sound and offers a much better prospect for accountability of the algorithmic systems there is one crucial aspect that it does not take sufficiently into account. Namely, a significant difference in speed of technology and ethical considerations and regulations. The speed of technological development is increasing it the context of the 4th industrial revolution and will only get faster with the prospect of quantum computers. In order to adapt mechanisms of human oversight have to go hand in hand with algorithmic development in the very process of design. Fab labs could be some of the potential places to look for inspiration for testing the algorithmic impact that is done by potential user groups together with diverse experts and researchers before it will be applied on a broader scale. In that way, potential harmful consequences could be prevented in the design phase adopting the trustworthy approach of responsibility and responsiveness to values and need of user groups and publics.

Blockchain technology might offer promise of enabling the processes trustworthy by design. Its potential lays in rebuilding public trust that is continuously declining both in the context of trust in institutions and increasingly declining trust in technology due to the questionable security of personal data and false information (Edelman, 2018). Both trends show that the strategy of promoting trust in institutions or in science and technologies (Smallman, 2018) has not been fruitful. On the other hand, the potential of Blockchain so far hasn't been sufficiently explored in terms of designing such a technology that would be trustworthy by design. Based on participatory and peer to peer approaches it could potentially ensure responsibility, accountability, and audit without outsourcing trust to centralized institutions or platforms.

Since Blockchain enables encoding the rules of trustworthiness in algorithms and software it offers the prospect of trustworthiness by design without requiring trust in central authority (Swan, 2015) or relying on centralized institution as a middleman, be it government or financial institution. Instead, it is based on the consensus algorithm thus introducing new management of trust through decentralized network and enabling building new social infrastructure (Vigna and Casey, 2018). By encoding participatory and co-designing processes directly in the design of the algorithm, it could potentially have an immense societal impact. It is basically an electronic ledger that can handle transactions of anything of value, from cryptocurrencies such as Bitcoin (Nakamoto, 2008) to smart contracts such as Ethereum (Buterin, 2013) or anything else that is of value.

Despite the growing interest in the blockchain technology (Suberg, 2016; Barker, 2015) there are still concerns about the huge amount of energy that it requires as well as security risks. Even before recent hack and theft from DAO (Decentralized Autonomous Organization), that happened in 2016, Eyal and Sirer (2013) warned about other ways in which the technology was vulnerable. They had shown how hackers and attackers can trick the consensus protocol so that the colluding group can take over control of the network mining power and compromise the system. But the development of this technology is still evolving both in terms of countering risks and also in using Machine Learning to make it more dynamic and adaptive (Hassan and Filippi, 2017) for encoding rules into code.

As analyzed in this chapter, new technological opportunities together with possible benefits also pose challenges and threat of exclusion, discriminatory practices, hindering diversity

and participatory processes since algorithms are not value-free. However, if it is openly acknowledged that algorithms are not value free and if the trustworthiness is encoded directly into code, the prospect of Blockchain technologies might be utilized in novel ways. There are already inspiring examples, such as developing a platform based on Ethereum for DAO that could be used as decentralized crowdfunding platform such as Betfynding (Jacynycz et al., 2016), or for encoding different governance models in DAO such as P2PMODELS (http://p2pmodels.eu/) that attempt to enable new ways for Collaborative Economy. Despite the obstacles in terms of energy, security or regulations Blockchain as open source technology gives an inspiring glimpse of the new prospects for trustworthiness and initiatives that might lead to new solutions and social infrastructure trustworthy by design.

1.4 Conclusion

Focusing on decision-making algorithms and their implications for the society we examined main ethical challenges as indicators that influence changes in trust. Case studies on racial and gender biases in decision-making processes powered by algorithms contributed to exposing value-leadenness of algorithms. Values are involved in developing or implementing algorithms, either entering into the design of algorithms from their designer, the institution or as emerging properties and due to technical restraints. Therefore, human prejudices and biases perpetuating in decision-making algorithms can cause unfair outcomes and discrimination.

However, in assigning the responsibility for the consequences of the decision-making algorithms, machine-learning algorithms pose an additional challenge. Simply assigning the responsibility to engineer or designer of algorithms, the users or to technology cannot be applied due to a specific amount of autonomy inherent to machine learning algorithms. The results of such learning process cannot be completely certain nor can unintended consequences be predicted, so that the algorithms can be a "black box" even to their designers. Furthermore, when it comes to responsibility and accountability of decisions that have been devised in this way it is also difficult to explain how decisions have been made. That makes decisions hard to understand or to challenge and complicates the questions of placing trust wisely.

In such intervened relation between human and algorithmic agents, it becomes even harder to assign responsibility and accountability and to determine if algorithms can be trusted. In determining who should be responsible, according to our analysis, the responsibility can be assigned only to human agents as they possess the aptitude for intentionality. Therefore we argue that human agents that possess capacity for intentionality and can be held responsible and an oversight over the process of the designing and implementing decision-making algorithms which makes them a potential subject of trust.

According to our analysis, algorithms on contrary cannot be trustworthy, as they do not hold the properties of intentionality. However, that does not exclude them from a moral agency, specifically in terms of autonomy of machine learning algorithms. As artificial moral agents, they nevertheless can be held accountable and thus contribute to the model of trustworthy algorithmic systems or organizations that are producing and implementing algorithms. Therefore, we argue that the problem of trust in digital technologies enabled by algorithms should be reframed as a

question of trustworthiness of algorithmic systems that joins human and artificial moral agency and related responsibility and accountability. It refers both to design and implementation in organizations that use decision-making algorithms - such as financial, medical institutions, police, law, etc., having an impact on the decisions that will affect significant aspects of our lives.

In the second part, we placed the discussion about benefits and downsides of regulating algorithms in the context of digital transformations. New digital transformations bring a wealth of opportunities in enhancing our lives, effectiveness, and efficiency and the prospect of using technological advances for healthy aging for example. However, new technological developments and automation will make lots of jobs obsolete and can have a huge impact on the widening inequalities gap, enhancing discrimination and jeopardizing fairness.

In this context, we were interested whether the regulation of algorithms would contribute to ensuring trust or it would rather stifle innovation instead. Analyzing the prospect of potential regulatory practices, we identified several challenges in its implementation. First relates to transparency that we understand as accessibility and comprehensibility of information that is hardly achievable in the context of algorithms.

There are several reasons why information would not be accessible but instead, be kept in secrecy. It might be for the national security, the privacy of data subjects, the autonomy of organizations or competitive advantages in companies that we closely examined within the recidivism bias example.

Comprehensibility is a second challenge within the quest for transparency. We call it a "black box" problem of machine learning algorithms because they are incomprehensible even to their designers. We examined this aspect of opacity more closely within the example of The EU General Data Protection Regulation (GDPR). It exposes the challenge of the comprehensibility requirement about exact working especially of machine learning algorithms because they can hardly be sufficiently explainable.

After examining the transparency challenges of algorithms, we conclude that even if the complete transparency might be achievable, it still does not ensure that the outcomes of the decision-making algorithms are fair or reversible or that responsibility is adequately taken into account. In trying to overcome accessibility and comprehensibility challenges we assessed three suggestions for regulation of algorithms. In the first proposal independent body or institution would serve as a regulator of algorithms, the second proposal is about the internal

regulating mechanisms enabled by algorithms and the third approach emphasizes the role of experts in the process of regulation.

According to our analysis, neither of the proposals for regulating algorithms pay sufficient attention to the aspects relevant for ensuring trustworthiness of algorithmic systems. In order to address the existing gap, we focus on the aspects of participatory processes in algorithms design. We further emphasize the importance of including in the process of algorithm design diverse actors who will be affected by the decisions.

And finally, we conclude the discussion on trustworthy algorithms by examining it in the context of the concept of trustworthiness. It shows that although algorithms could be held accountable, the question of their trustworthiness is misplaced, as they do not have the capacity for intentionality only inherent to human agents. We argue that trustworthiness should be related to the complex interrelation between human and artificial agents and we devise a model of trustworthiness that can be applied to different agents including organizations that are producing or using algorithmic systems.

In this context, we suggest the model of trustworthiness, which should be assessed based not only on competence and reliability but also in terms of responsiveness to potential concerns and values of trustees. We claim that is best done through co-creative processes early in the process where the design of algorithms takes place. Therefore we introduce a new model that is trustworthy by design that would enable participatory and inclusive processes. The application of the model is illustrated through blockchain technology that offers new ways of encoding the trustworthiness in the very design of the algorithms.

CHAPTER 2: TRUSTING INSTITUTIONS OR DECENTRALISED PLATFORMS?

2.1 Concepts of trust and trustworthiness

Within the current context of rapidly changing ways of life being introduced by digital transformations the question of trust gains renewed relevance. Furthermore, it will become even more pertinent as a source of social glue in this rapidly changing environment that some call 4th industrial revolution (Schwab, 2017). Although trusting relations have been attributed as crucial for social fabric, trust is in flux. Surveys repeatedly warn of growing mistrust in institutions (Edelman, 2017, 2018; European Commission, 2017). Since people are losing trust some line of research suggests that trust should not be relevant in institutions where it can be exchanged with cooperation (Cook et al., 2005). Optimism that new digital technologies offer in terms of distributed trust (Botsman, 2017) redefines the ways we place trust in individuals and complete strangers suggesting that trust is dramatically changing its form.

In order to tackle the assumption that trust in institutions is failing and that a new mode of trust in digital technologies is on the rise, we will take on the challenge of the very genesis of the debate on trust starting from the origins of the concepts of trust. Conducting the analysis of the current literature on the concepts of trust we aim at answering the main questions: Should we trust institutions when the stakes are high or decentralized platforms offer better prospect? Is trust emotional or we can base it on evidence? Could institutions or algorithms be trusted?

An attempt to conduct an analysis of the concept of trust and trustworthiness soon, however, has to diverge into an analysis of the plurality of concepts. There is no single concept of trust but rather simultaneously existing different concepts of trust and related trustworthiness. Those concepts are context related, often defending opposite arguments and are bolstered by examples covering very different topics, ranging from friendship, love, relations between mother and child to various institutional and historical settings.

In an attempt to categorize those diverse concepts of trust and trustworthiness first main distinction should be drawn between interpersonal and institutional concepts of trust. In the philosophical literature attention is mainly given to interpersonal modes of trusting relations (Baier, 1986; Pettit, 1995; Jones, 1996, 2012; McGeer, 2008), while second direction of research

on trust in institutions (Govier, 1997; Hardin, 2002; Potter, 2002; Townley and Garfield, 2013) has gained interest more recently.

The main focus of our research on trustworthy institutions will be built upon this second line of research that examines trust related to institutions. However, we first start with an analysis of various concepts of trust and trustworthiness, their definitions and distinct approaches in order to lay a basis of the state-of-the-art research on trust and trustworthiness. Apart from the first obvious distinction between interpersonal and institutional approaches to trust, there are also common features that can be found across different approaches. We identify them first as a three-part relation where A trusts B to do X (Baier, 1986; Holton, 1994; Hardin, 2002), secondly, as a prominent formulation of concepts of trust distinct from the concept of reliance (Baier, 1986; Holton, 1994; Blackburn, 1998; Walker, 2006), and finally the significance of trust that as such cannot be willed (Baier, 1986; Gambetta, 1988; Jones, 1996; Hardin 2002)

However, when it comes to motivational causes of trusting relations concepts tend to diverge significantly, from goodwill account (Baier, 1986; Jones, 1996), affective attitude (Jones, 1996, 2013), encapsulated interest (Hardin, 2002) to social norms (Dasgupta, 1988; Hardin, 2002; O'Neill, 2002a).

An overview of the main concepts of trust in both interpersonal and institutional setting (ranging from the trust as a cognitive approach to affective attitude, and based on various motivational causes) serves as a starting point for our critical analysis. Our hypothesis is that instead of trust rather trustworthiness should have a priority and even more so in an institutional setting. The main reason for that is that trust can be blind and misplacing trust in the one that is not worthy of it can lead to disappointment, disillusionment, even exploitation or misuse.

Therefore, we should be primarily concerned with the notion of trustworthiness of the trustee rather than initiating more trust per se regardless. Although the main focus in the literature review is on trust, several authors reversed the perspective recognizing the relevance of trustworthiness in interpersonal relations (Baier, 1986; Pettit, 1995; Jones, 1996, 2012; McGeer, 2008) or from the institutional or social perspective (Hardin, 1996, 2002; O'Neill, 2002a, 2013, 2014; Potter, 2002). We build on their work, extending it to the critical analysis of the models of trust that are not adequate for fostering trust in an institutional setting. Concludingly, we claim that models of trust are also far from being applicable in the trustworthy institutional design, therefore, proposing new trustworthy institutions model.

Our method of arranging different types of trust in a pluridimensional map gives us a new understanding and enables us to identify the missing links in each of the concepts. Although they are pertinent in the limited and specifically intended context and for a specific limited purpose, their application to the context of trustworthy institutions requires extension and reformulation. Based on the analysis of current concepts of trust we will propose three hypotheses:

1. Ever more prominent appearance and relevance of the strong thin mode of trust has yet failed to be acknowledged;
2. Affective and emotional aspect of trust should neither have a central role, or be excluded, but based on our elaboration of the unsustainability of knowledge deficit hypothesis, they should play a role next to cognitive aspect of trust; and
3. Instead of abandoning trust in institutions, we propose a different model of trustworthy institutions.

2.1.1 What concepts of trust have in common?

Apart from the first obvious distinction between interpersonal and institutional approaches to trust, there are also some common features that connect different approaches and can be found in several of them, both interpersonal and institutional. First, the formulation of concepts of trust is mainly defined in distinction to mere reliance that cannot be betrayed but only disappointed. Secondly, a prominent feature of trust is a three-part relation where "A trusts B to do X". And finally, mostly shared significance of trust is that as such it cannot be willed.

Those three basic characteristics are mostly shared across various conceptualizations of trust. However, when it comes to characterization of trust as cognitive or affective attitude and to defining the source of motivation of trust, the accounts are diverging across the spectrum. Our aim in this chapter is to distinguish basic characteristics of concepts of trust that are shared between three distinctive accounts of trust: goodwill account, trust as an affective attitude and encapsulated interest view.

The purpose of summarizing the main points of the currently existing accounts will then serve as a state of the art research on trust and a starting point for our analysis of aspects that can not be applied to the contemporary institutional context. It should further justify the need for the

new model of trustworthy institutions that will be developed in the next chapter. In what follows we will examine in turn each of those unifying characteristics of trust that we suggest are shared across different concepts.

First, trust is mostly formulated in its distinction from mere reliance. When trusting we are open and vulnerable to a possibility to be betrayed, a feature that is not present in mere reliance (Baier, 1986; Holton, 1994; Blackburn, 1998; Walker, 2006). This difference makes the cornerstone of the distinction between mere reliance and trust. It gathers the authors who criticize a purely cognitive approach to trust that is based on the ratio (Gambetta, 1988; Coleman, 1990). They claim that trust on contrary precedes having complete evidence and monitoring of its implementation and therefore entails the risk and exposure to the vulnerability. It also involves discretionary powers (Baier, 1986; Dasgupta, 1988) and therefore holds possible danger for the one who trusts.

The concept of reliance is mainly applied to machines, such as relying as on our watch to tell us the time. We can rely on inanimate objects and our reliance can be related to various aspects of "dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all" (Baier, 1986, p. 234). If we are relying on something we can be disappointed if the specific action is not performed, but we cannot be betrayed. Trust, on the other hand, can be directed only to things that have will.

A second common formulation of trust in philosophical literature, with slight modifications, is formulated as a three-part relation, where A trusts B to do X (Baier, 1986; Holton, 1994; Hardin, 2002). It is more specific than solely two-part relations of trust because it also articulates the domain of competence and commitment in which trust takes place. We can seldom trust someone generally in everything. We could trust scientists in his field of expertise, but not necessarily to conduct a surgery. Baier's three-part relation of trust is further specified within her entrusting model and articulated as "A trusts B with valued thing C" (Baier, 1986, p. 236). Thus, her account also includes the valued thing that is entrusted and lets another person take care of something we care about. Although there are several further variations of the formulation, the main structure is the same extending to related concept of three-place trustworthiness, as we will see in the next chapter, which starts exactly from this perspective of the trustworthiness.

The third conception of trust is that it cannot be willed (Baier, 1986; Gambetta, 1988; Jones, 1996; Hardin, 2002). As Baier clearly formulates "Trust me!" is for most of us an invitation which we cannot accept at will-either we do already trust the one who says it, in which case it serves at best as reassurance, or it is properly responded to with, "Why should and how can I, until I have cause to?" (Baier, 1986, p. 244). Jones further claims that trust cannot be willfully adopted because it is an affective attitude and can also not be demanded without grounds about goodwill, competence, and expectation that someone will be moved by our counting on them. However, she permits that we can decide that the evidence is enough to adopt attitude or beliefs, but we cannot do that regardless of evidence.

She specifies it as follows “While affective attitudes can't be willfully adopted in the teeth of evidence, once adopted they serve as a filter for how future evidence will be interpreted” (Jones, 1996, p.16). Although trust cannot be willfully adopted despite the evidence, it can be cultivated based on the grounds for trust. One of the main characteristics across concepts of trust is that it cannot be willed or demanded without evidence or grounds of others trustworthiness. This evidence can then imply competence and commitment, or further aspects such as goodwill, interest, moral integrity or other characteristics depending on the specific concept of trust.

Apart from those common grounds underlying concepts of trust, they significantly diverge when it comes to what motivates the one we trust or the one who is trusted to be trustworthy. Is trustworthiness motivated by goodwill (Baier, 1986; Jones, 1996), moral commitments (McLeod, 2002; Nikel, 2007; Cohen and Dienhart, 2013), based only on interests (Hardin, 2002) or social norms and constraints (Dasgupta, 1988; Hardin, 2002; O’Neill, 2002a).

Or could trustworthiness be motivationally neutral (Jones, 2012) as in “counting on” account for trustworthiness - according to which being moved by the fact that someone is counting offers unifying account of all different motivational modes. In the next section, we will examine three paradigmatic formulations of different motivations articulated within the goodwill account, trust as an affective attitude and trust as encapsulated interest. We will scrutinize them according to the criteria of adequacy for the institutional model of trustworthiness with a goal of singling out critical as well as potentially beneficial aspects.

2.2 Should we trust institutions when the stakes are high?

2.2.1 Goodwill account of trust

Should we trust juridical institutions to make fair decisions in predicting recidivism that will influence our jail sentence based on software that might show to be biased? Is it enough to rely on optimism of the goodwill and competence of institutions if life sentence might depend on their decisions? Questioning adequacy of the goodwill account of trust, when applied to the context of complex institutions, does not come as a surprise. However, addressing the challenge of the goodwill account of trust as being to narrow for the institutional context reveals another more relevant insight on the missing link of the strong thin mode of trust that we will tackle in more detail.

In interpersonal relations when we have to judge whether to place trust in another person when the stakes are high we could be assisted by evidence on character traits of the person based on close face to face relationships. But when we have to reflect on placing our trust in complex and distanced institutions such as medical services or jail sentence when the stakes might be even higher, how should we place our trust wisely? In those situations related to trust in complex institutions, it is crucial to recognize that another mode of trust is at place, namely strong thin mode of trust, the concept best elaborated by Hosking (2014). In order to address those various modes of trust and identify the missing link in the context of trust in institutions, we will elaborate on the goodwill account of trust that primarily addresses interpersonal relations. Then in the second part, we will place this account of trust in the broader context of trust in institutions, focusing on the strong thin mode of trust.

Our analysis of the goodwill account of trust will first focus on the model developed by Annette Baier (1986, 1991, 2013) as her conceptualization of trust was to a great extent a starting point for the further interest in that topic. Then, we further elaborate on the concept of goodwill account of trust significantly extended in the early work of Karen Jones (1996). Although there are some differences between them, their joint focus is on interpersonal relations rather than trust in institutions. Similar to both accounts is also understanding of trust as reliance or attitude of optimism of trustee's goodwill and competence. The goodwill is perceived as the main motivational aspect according to this account, claiming that goodwill is infused to the one who is

trusted by the very act of trusting (Jones, 2013). After outlining the main characteristics of interpersonal goodwill account in both works of Baier and Jones, we will then examine aspects of the goodwill account that could be problematic in the context of trustworthy institutions.

First, shared characteristics of the goodwill account of trust can primarily be understood as interpersonal trust. It can be applied to relationships between individuals, mother and child, lovers and furthermore "it is the trust always found in friendship, often found between professionals and their clients, sometimes found between strangers, and sometimes even, between people and their governments" (Jones, 1996, p. 5). In this broader sense interpersonal trust thus stretches not only to the relations between individuals but also relate to institutions such as governments. That aspect of interpersonal trust will prove to be beneficial in our attempt to determine whether the account can adequately be applied also in the relation between publics and institutions. Second, shared characteristic of the goodwill account of trust refers to both goodwill and competence, as goodwill account of trust is grounded on reliance (Baier, 1986) or attitude of optimism (Jones, 1996) of trustee's goodwill and competence. Apart from these two shared characteristics, both Baier and Jones further develop the concepts of goodwill account of trust in distinctive ways.

The significance of Baier's concept is that she develops entrusting model of trust as a three-part relation including the valued thing that is entrusted "A trusts B with valued thing C" (Baier, 1986, p. 236). In leaving discretionary power to the trustee in taking care of the entrusted thing, the trustor accepts the state of vulnerability and potential betrayal. This can also be seen as the main distinction from mere reliance, which cannot be betrayed. The goodwill account of trust developed by Baier has sparked further interest in that topic and resulted in different variations of the concepts that followed among which is the affective account of trust developed by Karen Jones.

Karen Jones' (Jones, 1996) articulation of the goodwill account of trust extends the original concept in two main directions, in terms of affective attitude and expectation. She describes affective part of her concept of trust as an affective attitude of optimism about a trustee's goodwill and competence. Cognitive criteria of her concept of trust, on the other hand, is within the expectation that trustee will be favorably moved by the realization that we are counting on her. According to Jones's account "trust is an attitude of optimism that the goodwill and competence of another will extend to cover the domain of our interaction with her, together with

the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her" (Jones, 1996, p. 4).

Another distinctive characteristic in Jones' work (Jones, 2012, 2013, 2017) is assigning neutral motivation to being trustworthy in her "counting on" account that gives a clue about the reasons to be trustworthy. According to this account, the motivation to be trustworthy lies in the very fact that others are counting on the trustee. In introducing this neutral motivational framework, she hopes to accommodate various other motivational impulses. This new reformulation of her account of trust brings a welcome twist from the purely goodwill account. She further extends her new and reconceptualized account of trust by including the notion of responsiveness, that will in its refined definition also play a crucial role in devising our model of trustworthy institutions later on.

2.2.2 Strong thin mode of trust

As we have seen, the goodwill account of trust is based on the reliance on goodwill and competence of trustee and as such is also primarily used in the interpersonal context. Its potential inadequacy in institutional context and criticism of being too narrow should then not come as a surprise if we consider it's mainly interpersonal character. However, conducting the exercise in identifying critical points of this first concept of goodwill account of trust enables us to find the missing link when understanding the new emerging and prevalent modes of trust. By applying the concept of goodwill account of trust to the institutional setting we uncover the missing link of the strong thin mode of trust (Hosking, 2014), which crucial relevance is being overlooked. Our analysis will start by tackling (1) the lack of clear definition of goodwill; (2) too narrow aspects of account to be applied to institutions, and finally; (3) the relevance of the strong thin mode of trust that fails to be acknowledged.

The first general observation and critic of the goodwill account of trust is its imprecise definition of goodwill. It is hard to understand what exactly is meant by goodwill as neither Baier nor Jones clearly defines it and the term is understood very broadly in their formulations of goodwill accounts of trust. Jones formulates the goodwill as follows:

> "There are a number of reasons why we might think that a person will have and display goodwill in the domain of our interaction with her. Perhaps she harbors friendly feelings towards us; in that case, the goodwill is grounded on personal liking. Or perhaps she is generally benevolent, or honest, or conscientious, and so on" (Jones, 1996, p. 7).

The meaning of goodwill in this sense can be understood in a very broad way ranging from friendliness, liking, honesty to benevolence and does not offer a very precise account of this basic term of the goodwill account of trust.

In her later work Jones (2012, 2013, 2017) introduces changes in her goodwill account of trust and also recognizes the missing clarity in defining what goodwill exactly refers to. Furthermore, at that point, she also brings an interesting twist no longer separating goodwill from responsiveness. Her reformulated concept of trust can now be understood as follows: "A trusts B in domain of interaction D if and only if, A has an attitude of optimism that B's competence and responsiveness to the fact that she is being counted on will extend to cover that domain" (Jones, 2017, p. 15).

Apart from that general observation of imprecise definition, there are further points of the concept of goodwill account of trust that could be criticized if the account is to be applied to an institutional setting. The most prominent of them is that the concept of goodwill is too narrow to be applied to institutions (Blackburn, 1998; Walker, 2006). The critic of being too narrow to be applied to institutions nevertheless does not change its justification in the context of interpersonal relations, such as for instance between mother and child or friends where it is mainly applied.

However, there are several further lines of argument that prove goodwill account inadequate when applied to the institutional setting. The first reason for its inadequacy, as we have previously noted, is that goodwill can easily be replaced with other motivations for trustworthiness. Those motivations could be based on goodwill, but also either or moral commitments (McLeod, 2002; Nikel, 2007; Cohen and Dienhart, 2013), on interests (Hardin, 2002) or social norms and constraints (Dasgupta, 1988; Hardin, 2002; O'Neill, 2002a).

The second reason, of goodwill inadequacy, can be inferred from Jones' articulation of the "counting on" concept of trust according to which goodwill is not satisfactory without the additional expectation that the trustee will be favorably moved by the fact that we are counting on them (Jones, 1996). Or in the later modification of Jones' goodwill account of trust, goodwill is not even separated but only a part of responsiveness (Jones, 2017).

The third reason that can be related to its inadequacy is the obvious realization that caring cannot always be relevant in large institutional settings or in the encounters with strangers. Fourth, the inadequacy of the goodwill account in the large institutions could also be criticized as imposing an unnecessary distinction between trust and reliance. The very core of the goodwill account of trust is based on the distinction between trust and reliance, whereas that distinction loses its necessity in the institutional setting as Hawley (2017) elaborates in her argument against separation of trustworthiness and reliability.

And finally, we claim that the attempt to apply interpersonal to the institutional type of trust is unsuccessful because the relevance of the strong thin mode of trust had failed to be recognized. Therefore, we suggest to draw back on the differentiation between modes of trust best formulated by Hosking (2014) from mostly used and understood modes such as strong thick and weak thin to not so well understood and overlooked strong thin mode of trust. Usual modes are strong thick and weak thin modes of trust. Strong thin mode of trust can be present in highly valued interpersonal relationships where we can have extensive knowledge about the person that we trust, whereas weak thin mode of trust would be mostly related to institutions about which we are not very well informed, but where we also do not commit great resources to those institutions. The significance of the strong thin mode of trust, on the other hand, is this risky combination of our high stakes that are involved, such as health, life, jobs, while at the same time not having means and extensive knowledge on the complex institutions that would enable us to place trust wisely. This strong thin mode of trust is mostly not acknowledged, although it plays a central role in the context of a possible application of goodwill account of trust to institutions.

Hosking (2014) introduces four main modes of trust: strong thick trust, strong thin trust, weak thick trust and weak thin trust. The first general term is strong trust that he illustrates is used

> "for relationships to which individuals commit valued resources—which may be the preservation of their health, beliefs, customs, home, and way of life, their profession or job, provision for their children or their own old age. That would include trusting the quality of education in a school or college, taking out a mortgage to buy a house, committing one's free time to voluntary work for a charity or religious movement, or placing savings in a pension scheme which may not bring any benefit for several decades

> but could prove a godsend in old age. In all these cases, decisions will normally be preceded by serious reflection and the weighing up of alternatives" (Hosking, 2014, p. 33, chapter 3).

In the case of strong trust, the stakes and our committed values are high opposite to the weak mode of trust.

Another distinction between thick and thin modes of trust can be understood as follows: "'Thick' trust rests on extensive knowledge, resulting from frequent or close contact with the person or institution one trusts, whereas 'thin' trust is based on slight knowledge, on infrequent or superficial contact" (Hosking, 2014, p. 33, chapter 3).

Presumably, he does not propose a strict borderline between those concepts but a rather slow gradation as well as their different combinations. For example "strong thick trust might be getting married; of weak thin trust buying food in a supermarket" (Hosking, 2014, p. 34, chapter 3). Weak thick trust suggests that "one does not always need to risk major resources in a close or frequent relationship; one risks little in lending a close colleague the bus fare home" (Hosking, 2014, p. 34, chapter 3).

Personal contacts and living in small communities provide us with an abundance of evidence of past experience and character traits making it easier to predict future actions and judge if we should place our trust or not in the strong thick mode of trust. However, in the strong thin trust that is expanding for last half century, as Hosking (2014) suggests, the radius where we place trust in ever bigger and more complex organizations and institutions hinder our ability to judge where to place our trust. We do not and cannot handle the necessary amount of evidence to place trust wisely in the same way it can be done in strong thick modes of trust.

The most important insight of Hosking categorization of different modes of trust is his argument that strong thin trust "is ever more prevalent in our social life today, as a result of cumulative changes which have been taking place at least in the West for a very long time. One of the main reasons we often misrecognize trust today is that we have not been aware of the growing predominance of strong thin trust" (Hosking, 2014, p. 34, chapter 3).

Conceptualized in this way in a broader context of different modes of trusting relations from strong, weak, thin, thick trust and their variations we can better understand the new strong thin mode of trust that is gaining influence in contemporary societies. The first logical implication is that concepts of trust that we have been analyzing and that are mostly present in interpersonal

relations would be formulated within strong thick modes of trust that cannot be adequate for the new types of trusting relationships in the large complex institutions where a strong thin mode of trust is predominant. Different modes of trust are dependent on how high are both risks involved and the values it holds for us, as well as the amount of relevant information and evidence that we can acquire.

To clarify this distinction we have to look closer to the possibility of trust in larger institutional settings. Here the benefits of face-to-face relations characteristic for the thick mode of trust are not accessible. Furthermore, contrary to interpersonal relations our knowledge of complex and remote institutions cannot be so extensive, particularly in terms of all different specializations and expertise that might be relevant even in one specific case. Although the thick mode of trust does not apply here, the strong mode of trust can still be prevalent.

In different contexts of large institutions, the stakes might also be high, our life can depend on medical institutions or scientific medical institutions research results pointing again to the strong thin mode of trust. According to Hosking this mode of trust is more present today than we acknowledge: “One of the main reasons we often misrecognize trust today is that we have not been aware of the growing predominance of strong thin trust” (Hosking, 2014, p. 34). Whether the dominance of the strong thin mode of trust in institutions happens in parallel or induces further exploration of new modes of trust in digital platforms yet remains to be examined.

2.3 Is trust emotional?

2.3.1 Trust as an affective attitude

The previous example of personalization algorithms filtering relevant content instead of us might make us more efficient, but also could contribute to creating an echo chamber by reinforcing our pre-existing beliefs. Can we then claim that we place our trust in a reflective way or in an emotional way led by the sources and people we like. Could trust be emotional and resistant to evidence and how important is our trust in the source of information provider. In order to address those questions, we will first analyze the concept of trust as an affective attitude in an attempt to understand the emotional character of trust and how beliefs resistant to evidence are being formed.

This second prominent concept of trust is trust as an affective attitude. Common ground with previously examined goodwill account is their shared *attitudinal* character, or more precisely they are both accounts of trust as an *attitude* towards trustors' goodwill and competence. Both Baier (1986, 1991) and Jones (1996) in their understanding of trust have attitudinal aspect as a common trait, although they differ in its further articulation. While Baier elaborates on an entrusting model, Jones focuses on the affective attitude model of trust. When examining the concept of trust as an affective attitude we primarily have to focus on its first formulation in the work of Karen Jones.

In Jones' account of trust as an affective attitude, there are two essential parts, cognitive element and affective element to which Jones (1996, 2013) assigns the central role. Jones' emphasis on the *expectation* that the trustee will be favorably moved by the realization that we are counting on her is presented as the cognitive element of the concept, but the emphasis on the *affective attitude* of optimism about trustee's goodwill and competence still holds the central place in Jones' concept. Our analysis of two constitutive elements of Jones's concept will simultaneously attempt at exposing its critical aspect in the context of application in institutions. We will claim that the affective aspect of trust, although relevant and necessary to be acknowledged, nevertheless cannot have a central role but should rather be in interplay with cognitive aspect and centered around the priority of trustworthiness. But first, we start with the analysis of the cognitive and affective aspect of Jones' account of trust as an affective attitude.

First, the cognitive element of trusting relation is tied to the expectation that the one who is trusted will be moved by the thought that someone is counting on them. If this expectation is not fulfilled and if the person is not moved but rather only reliably benevolent, Jones would consider that person only reliable but not trustworthy. In that way, she articulates the distinction of trust from reliance and trustworthiness from reliability. Further elaboration of the cognitive element of Jones account suggests that trust has to be grounded in the evidence about trustee's competence, goodwill and - in her later articulation – responsiveness to our counting on them. This aspect of affective account of trust that implies the necessity of evidence for trust to be well grounded, in Jones' account, however, does not seem to have a dominant role.

Both Jones (1996) and Baier (1886) claim that trust as an attitude cannot be willfully adopted or demanded. It has to be grounded on the goodwill, competence, and expectation that someone will be moved by our counting on them. As Jones articulates:

> "Because trust involves an affective attitude, it is not something that one can adopt at will: while one can trust wisely or foolishly, trust cannot be demanded in the absence of grounds for supposing that the person in question has goodwill and competence and will be likely to take into account the fact that one is counting on them. This is not to say that there can never be an element of decision in adopting beliefs or attitudes. We can, for example, decide that the evidence we now have is enough to support the belief, but we can't just decide to believe regardless of the evidence. While trust cannot be willed, it can be cultivated. We cultivate trust by a selective focus of attention toward the grounds for trust and away from the grounds for distrust" (Jones, 1996, p. 16).

Although Jones suggests that trust has to be grounded in the evidence about trustee's competence, goodwill, and responsiveness to our counting on them, nevertheless she does not assign a dominant role to this aspect. Jones does not regard trust primarily as a belief about others trustworthiness, but she rather bases it on the centrality of affects and emotions.

That brings us to the second element of trust as an affective attitude, the aspect to which Jones assigns a central role. Our analysis in the context of a possible application of the account of affective attitude to the institutional trustworthiness model will therefore mainly be focused on this affective aspect, arguing that the adequacy of its central role, however, is not so obvious. The argument on the central role of affective attitude can be understood within two main points. First, trust is possible without having a belief about trustworthiness illustrated by Jones' examples of

cultivating trust or relying on intuitive assessment. Here we will argue that this claim does not seem obvious when applied to a broader context of institutions.

A second point, however, is elaborated as trust that can give rise to beliefs resistant to evidence. It will on contrary provide beneficial insights for the future articulation of our model of scientific institutions. We will present in turn those two points of Jones' argument that propose centrality of affective attitude. Then we will analyze each of them through the lens of their adherence and applicability in the context of potential institutional trustworthiness.

The first argument that suggests the centrality of the affective aspect of trust in Jones' account is that trust and distrust are primarily not beliefs and that trust can also be justified without having a belief about the trustworthiness of another person. Jones illustrates this claim with two examples where trust leaps ahead of evidence and is governed either by forward looking or backward looking considerations (Jones, 1996).

It can either cultivate trust towards a certain group of people by focusing attention and an interpretation of the aspects that are supporting our trust in this forward-looking mode of trust leaping ahead of the evidence. Or considering backward looking consideration, evidence might be surpassed by our intuitive assessment. In elaborating both of those considerations Jones uses interpersonal examples. Cultivating trust can sometimes induce trustworthy behavior when applied, as in Jones' example, to relation to child or partner. It can also be applied in intuitive assessment, as in her illustration of the personal attitude towards a salesman, whom we might find suspicious based on our intuitive reaction.

Despite positive recommendations about the salesman that we might have had received from our friend we might still hesitate in trusting this salesman although we cannot provide justification for our suspicion. Here, our intuition has primary say and the affective aspect of trusting relation is predominant. We do not dispute that these two examples provide possible situations of trust leaping before evidence, and thus might be used to support the argument for the centrality of affective attitude of trust. However, their relevance explained in Jones examples is mainly in interpersonal relations of trust, but would not be adequate incentives for its application in the institutional setting and for that purpose other mechanisms would have to be developed.

The second argument in favor of the centrality of affective attitude claims that trust gives rise to beliefs resistant to evidence. It, therefore, implies patterns of interpretation and seeking out evidence that is self-confirming and centered in the affective attitude. Jones admits, however, that

trust that gives rise to beliefs resistant to evidence is not limitless. We trust someone and resist evidence that proves them guilty, but if there is enough evidence that they are guilty we might not trust their innocence anymore.

This point that our beliefs can be resistant to evidence or seek the evidence that would confirm our pre-held beliefs brings important implications that we will further elaborate in the context of trusting scientific institutions and hesitance in accepting the claims based on evidence. This emotional aspect of our trusting attitude is not to be disregarded because, as we will show, it might prove to be beneficial also in the institutional context.

> In this regard, Jones further clarifies the central role of affect for trust arguing that "the attitude of optimism is to be cashed out not primarily in terms of beliefs about the other's trustworthiness, but rather-in accordance with certain contemporary accounts of the emotions -in terms of a distinctive, and effectively loaded, way of seeing the one trusted. This way of seeing the other, with its constitutive patterns of attention and tendencies of interpretation, explains the willingness of trusters to let those trusted get dangerously near the things they care about" (Jones, 1996, p. 4).

The way of seeing the other or perspective that we will have when interpreting their actions and motivations, according to Jones' account, will thus be different depending if we trust or distrust the other. Therefore, if we already trust someone we would interpret their motivations and actions to be self-confirming of our already existing trust. But if we do not trust the other we could interpret the same actions completely different, focusing on different arguments that would confirm our distrust.

As previously elaborated, articulating the account of trusting wisely seems to be of utmost importance because the vulnerability of trusting relation can pose serious harms to individuals and publics in their relation with complex institutions when the thin strong mode of trust is at place. But it has to be admitted that trust sometimes is based on the beliefs resistant to evidence as experimental psychology (Kahneman, 2013) articulates in great detail. That point should surely not be overlooked, although it does *not imply* that emotional and affective attitude should have priority in the institutional trustworthiness.

2.3.2 Beliefs resistant to evidence

In both of Jones' arguments for the centrality of affective aspect of trust, there are crucial insights into the workings of the beliefs that can be resistant to evidence or seek the evidence that would confirm our pre-held beliefs. The existence of such predominant modes of affective attitudes could be observed not only in interpersonal relations but also in institutional and other social settings and should therefore not be neglected or overlooked. In the final instance, however, we claim that Jones' proposed centrality of the affective attitude has to be rejected and we show it inadequate within the model of trustworthy institutions.

Nevertheless, although not taking the predominant role, we emphasize that the workings of the affective and emotional attitudes have to be adequately recognized. That aspect will also prove to be of particular importance in the context of scientific institutions and related to trust in expertise examined in the final chapter. At this point, we will illustrate the importance of affective attitude in the context of science, although rejecting its central role in the institutional setting.

Apart from interpersonal examples that Jones (1996) provides in institutional context it is also possible to observe examples of trust leaping in front of evidence, interpretation of aspects that are supporting our trust and potentially even surpassing evidence. We will illustrate it with the example from the scientific institutions from the second chapter of this thesis on trustworthy scientific institutions. When it comes to trusting scientific claims, it can often be closely related not so much to extend and quality of evidence, as to the question if we already believe the source that is sharing the message.

As research in experimental psychology (Kahneman, 2013) suggests, very often the way we place our trust is not reflective. People often rely on their System 1 that is intuitive, emotional and fast in drawing conclusions, but also prone to error and biases. The impact of System 1 is that we often trust people and sources that we like, often regardless of more evidence that might be provided that would suggest the contrary conclusion, although to some critical point as Jones explains.

Dan Kahan's research on ideology, motivated reasoning and cognitive reflection (Kahan, 2013, 2017) further suggests that our trust in scientific claims is often closely related to our trust in the source of information depending if it corresponds to our beliefs and assumptions where the

exact expertise does not have to be crucial. In the same vein of research, Kahan shows that public understanding of science does not have to relate to its acceptance. When asked about science claims in the contested subjects such as climate change or evolution, members of different ideological groups have a very similar level of understanding. But when the question is framed in terms of their personal beliefs, then answers tend to diverge much more. This effect of protecting the views of their group or community is also further strengthened by the evidence of cognitive science elaborated in the work of Steven Sloman and Philip Fernbach on knowledge illusions (Sloman and Fernbach, 2017). In various examples, their research results further elaborate workings of System 1 and employment of emotions when it comes to protecting our communities of similar beliefs and attitudes.

Another line of research that relates to the emotional aspects and pre-existing beliefs and values closely related to public trust in science is the example of vaccine hesitancy (Goldenberg, 2016). Her research shows that knowledge deficit argument cannot be sustained, as more scientific evidence does not necessarily change public's beliefs. Their preconceived beliefs seem to be resistant to evidence, trust cannot be established and further evidence that is confirming their belief that vaccine is not safe is sought for self-confirming belief. In this case statement "Trust us we are experts" also does not seem to work. It becomes obvious that the cognitive aspect in trusting relation that is based on providing evidence for trust is not the only one to take into account. The public's values and concerns, as shown in the Goldenberg case study can be valuable guidance on the further aspects of vaccine studies that are of public concern but have not been taken on board in the research agenda. What becomes clear in this case is that the public does not necessarily lack the understanding of the scientific issue and evidence at hand, but rather does not accept the values that are underlying scientific research.

If the affective public's concerns and values would be taken on board and addressed including in deciding on the research agenda, then the public might develop a more trustful relation. The trustworthiness of scientific institutions that take into consideration the public's concern would be regarded as one in which it is wise to place trust. Current assurances of the trustworthiness of scientific experts and institutions are not always sufficient for publics to place trust, the sentence "Trust me I am expert" is not enough to ensure trust. Therefore, it should be emphasized that although cognitive judgement should play a central role in the trustworthy institutional design, the reality of the emotional aspects of the trusting relation should not be

disregarded. Affective aspects of trust should also be taken into account, either in form of the public's concerns and values, as we have just elaborated, or in other forms that are prone to possible manipulation and misuse evident in various examples of System 1 mechanisms. Both aspects would contribute to the possibility of placing trust wisely when it comes to the relationship between publics and trustworthy institutions.

Concluding, it should be realized a significant role that System 1 can have on placing our trust unreflectively, be it in interpersonal or institutional settings. However, that still does not presume the central role of affective attitude as Jones suggests. On, contrary, due to the reality of tendency to rely on System 1, when it comes to important decisions as Kahneman (2013) suggests we have to slow down and also employ System 2 in which science mostly operates. Since communication can fail due to the emotional aspects in order to place our trust intelligently epistemic and ethical norms are needed in order to clarify the steps needed to correct prejudices (O'Neill, 2013). The implications for the prospect of trustworthy institutions, therefore, requires the institutional design that enables overcoming the solely functioning of System 1, while at the same time developing the ways of acknowledging it and talking to it. Otherwise, those Systems could easily diverge into two distinct directions.

We do not dispute those emotional aspects of trust interplay with cognitive aspects, what we take issue with in the account of interpersonal trust is the priority given to trust as an affective attitude and the predominant role that emotions take in the interpersonal account of trust. Priority of emotional responses leading our trusting relations in an institutional setting would hinder us in placing our trust well. Our intuitive reaction about an individual in a specific institution could not be an adequate reason to judge institutional trustworthiness. Cultivating our trust can also not be taken as an argument for inducing trustworthiness on behalf of the institution. Rather than being beneficial, those emotional aspects of trust can potentially rather have harmful consequences for the publics and misleading influence in our interactions with powerful institutions. Instead of enabling us to judge the trustworthiness of institutions in order to place our trust wisely, encouragement to trust without having beliefs about trustworthiness can have serious consequences of misplaced trust in the institutional context.

Therefore we argue for developing a model of trustworthy institutions with a priority of cognitive aspect of trusting relation without denying the emotional aspect of trust. A more adequate concept would, therefore, be to adopt beliefs of trust based on sufficient evidence on the

aspects proposed by O'Neill (2002b), aspects that judge competence, honesty, and reliability. Being able to assess relevant properties of another person's character traits or institutional trustworthiness would contribute to placing our trust wisely. The emotional aspect, in this sense, should be taken into account, but in interplay with the cognitive aspect of trust based on the belief about other's trustworthiness.

2.4 Abandoning trust in institutions?

2.4.1 Trust as an encapsulated interest or social norms

And finally, in addressing the question whether trust in institutions should be abandoned and whether cooperation is possible without trust we will examine the third concept of trust as encapsulated interest. Hardin (2002, 2006) has introduced the concept in which motivation to be trustworthy is based on interest. According to this view, the trustee's interest motivates him to be trustworthy and encapsulates the interest of the trustor mostly to maintain the relationship.

In conducting our analyzing of the encapsulated interest concept, we will first examine the attainability of interest being its main motivational drive that in the context of institutions might prove not to be the most adequate explanation. Our further analysis of other aspects of Hardin's proposed encapsulated interest view will be conducted from the trustworthiness perspective which should result in mapping different possible incentives for trustworthiness. At the final instance, we will focus on Hardin's later work (Cook et al., 2005) examining his proposal of cooperation that could be possible without trust, and therefore making trust obsolete in institutions and organizations. We will analyze his proposal within the framework of the model of trustworthy institutions.

At first, it might seem that the concept of trust as an encapsulated interest would have a more promising role within the context of trustworthy institutions than the goodwill account of trust or trust as an affective attitude view. It is because, unlike the other two interpersonal concepts of trust it presumes that interest as motivation can be applied not only to individual relations, but also to social systems and institutions. Apart from this distinction between interpersonal and institutional modes of trust, all three concepts share the same idea that we can not choose to trust or that trust can be willed because it is cognitive notion based on knowledge about the trustee.

The main idea of the concept of trust as encapsulated interest Hardin explains as follows "You can more confidently trust me if you know that my interest will induce me to live up to your expectations. Your trust is your expectation that my interest encapsulates yours" (Hardin, 2002, p. 5). On another hand, if we have the wrong information about one's interest than trust can be mistaken. Considering the application of this model to the trustworthy institutions there are several points to be noted. First, taking interest as a motivational drive for trustworthiness overlooks other possible incentives that might be taken into consideration in institutional design, such as responsibility towards trustor or the publics. Second, in a complex institutional settings, it might also be more difficult to have enough evidence and knowledge about the one we trust because we cannot rely on face to face personal relationships that are available in the interpersonal relations. Thirdly, as trustor is mainly focused on his own interest that is encapsulated by the trustee in a specific domain of interaction, he might not be aware of another trustee's interest that could trump his own. This is the risk Hardin (2006) also acknowledges "there is some risk that my interests will trump yours and that I will therefore not fulfill your trust in me" (p. 19).

Furthermore, although the encapsulated interest view Hardin (2002) formulates as interest and not self-interest, in several practical examples it can easily be transformed into self-interest and selfishness. The interest and incentive to maintain a relationship in the institutional setting is explained as threefold. First, because this relationship is valuable to trustee, or secondly because he does not want to risk the reputation of being untrustworthy. The third aspect is evident in interpersonal relations that are motivated by friendship and/or love that gives the incentive to encapsulate one's interest is valuing values of others well-being.

Considering this expose it seems that interest does not necessarily have to be self-interest but could also relate to well-being of the other, integrating within it the affective side as well. Nevertheless, as this third aspect relates to the well-being of the other and would mostly be applied to friendships and towards people we love, it still remains only within interpersonal relations. In all other cases including the institutional context we are interested in, it is still self-interest that would have a predominant role. Furthermore, it is not totally clear what exactly is meant by "interest", therefore it can, in fact, be understood as self-interest if trustees other interest trumps the interest of the trustor. We do not have to look far to find examples of the devastating

consequences of selfish conduct in the institutions to find various cases in financial institutions for instance.

If we consider those critical points in the context of the application of the encapsulated interest concept of trust to trustworthy institutions, interest may not necessarily be the kind of motivation we would want to incentivize in the institutional design of institutions. One of the prominent examples could be within financial institutions, where the pursuit of interest showed to be easily turned to selfishness resulting in misuse of trust that the public placed in institutions. Here, also the trustee's interest or the interest of the financial institutions can trump the interest of the trustor and again indicate that the account of encapsulated interest view based on the interest as a motivation of trustworthiness should be replaced with more adequate motivational incentives.

Concludingly, it seems that neither motivation for trustworthiness that lies in interest (often turned into self-interest) or on goodwill, could offer the prospect of incentives for trustworthiness that would be adequate assurance of the prospect of placing our trust wisely. Therefore, in the next chapter, we will attempt in formulating more effective prospect for the trustworthy institutional design that would enable responsible and responsive conduct in the institutional setting.

2.4.2 Cooperation without trust

In this section, we analyze Hardin's proposal that trust is not attainable within some institutional contexts and his suggestion of cooperation within institutions that is possible without trust (Cook et al., 2005). Contrasting this claim to his earlier formulations of "creating institutions that help secure trustworthiness thus help to support or induce trust" (Hardin, 2002, p. 30) results in conflicting view on the role of trust in institutions from irrelevant to the institutions that induce trust that needs to be looked at more closely.

In his view on encapsulating interest, he formulates the argument for replacing trust with different mechanisms in the institutions due presence of asymmetric power relations. The encapsulated interest view is thus limited only to symmetric relations where dependency from both sides is at place; it includes the trustor who expects his interest to be encapsulated and the trustee who has an interest in maintaining relations. If the later cannot be ensured in the

asymmetrical power relations, and both parties cannot hold their commitment to the trusting relation, then according to this view, trust has to be replaced with different mechanisms.

Trust is therefore related to power relations between trustor and trustee (Hardin, 2002, 2006; Cook et al., 2005) and furthermore, in the institutions that enable cooperation, there is no need for trust (Cook et al., 2005). It might additionally be presumed that it would be difficult for the publics to judge the competence and motivation of institutions in all relevant domains of interaction. Therefore it is hard to judge the trustworthiness of powerful institutions and to trust them, and it is furthermore not even achievable where power inequality exists.

One of the possible examples is the relation to governmental institutions, which we cannot choose to substitute with another, so we must cooperate. In this context of power relation with institutions that have power over us, according to encapsulated interest view trust cannot be ensured. Or formulated further,

> "the fact that we are forced to deal with this single agency for certain important matters means that it has enormous power over us. We do not have even vaguely similar power over it, so that we suffer from power dependence. This means that we cannot trust these powerful institutions, or that at least the possibility of trusting them is severely undercut, especially in the encapsulated interest sense of trust, because my power dependence undermines any hope I might have to get you to reciprocally cooperate with me" (Hardin, 2006, p. 152).

The significance of the power relations specifically in an asymmetrical constellation of power dependence brings a new light to the trusting relations rightly questioning the capacity to trust. Numerous current studies and surveys listed in the introduction can only confirm this change in trusting relations in institutions. This view does not envisage possible mechanisms of the institutional design that would restrain power or incentivize other aspects of conduct that might enable trust in the trustworthy. Nevertheless, it can serve as a starting point in replacing trust in institutions with trust in digital platforms and incentivizing prospects of new technologies such as blockchain to explore other decentralized power models that would enable trust.

Hardin's concept of encapsulated interest should also make it easier to assess interests that are guiding our relations than other motivational aspects originating in moral, normative or other possible incentives. However, this view seems to present only one aspect of internal motivations for trustworthiness, next to external and mixed inducements. Those other possible ways of

ensuring trustworthiness “compelled by the force of norms” (Hardin, 2002 p. 52) are elaborated in various accounts of trust (Dasgupta, 1988; Hardin, 2002; O’Neill, 2002a, 2002b) and are motivationally neutral in contrast to goodwill account or encapsulated interest view for instance.

In describing trustworthiness as our commitment to fulfill the trust that someone put in us, “trustworthiness is a motivation or set of motivations for acting" (Hardin, 2002, p. 31), Hardin proposes three categories of reasons to fulfill a commitment of another's trust in us. They can either be internal, external or mixed inducements. First, internal inducements include encapsulated interest “we might simply have or somehow adopt a relevant disposition. We can do this by bootstrapping a bald commitment, from moral compunction, or out of character or habit” (Hardin, 2002, p 29). Apart from those internal inducements to which we have given more attention in the previous elaboration of concepts of trust, there are two other categories that are dependent on social control and norms. Hardin understands them from the perspective of our long-term interests and as mechanisms that can often conflict with our interests. Second, as external inducements, we can refer to law, legal regulations or other institutional constraints

> “we can subject ourselves to external inducements. We can try to arrange matters so that our interests will be aligned with our commitment when the time for acting on that commitment comes. We can do this by setting up ad hoc personal devices or by relying on societal and institutional devices. An individual cannot do much about the availability of societal and institutional arrangements – they either exist or they do not. But if they do, I can hope and often even contrive that they will help me rely on you. Legal and other institutional constraints can give us strong incentives to be trustworthy” (Hardin, 2002, p. 29-30).

The final, third category of norms can be understood as mixed inducement where

> “we can subject ourselves to a mixture of internal and external inducements. We can be induced by norms that motivate and even sanction behavior. Norms can evidently be internalized, so that we simply act from them without need of sanction. (…) I discuss norms primarily in their external role, in which their effectiveness depends on the success of sanctions from others” (Hardin, 2002, p. 29).

Those mixed inducements are a combination of internal and external inducements such as norms that can also be internalized in a way that we act upon them even though we will not be sanctioned. Norms might be very useful tools in guiding behavior in institutions and could be an

integral part of the institutional design aimed at enabling trustworthiness in the institutions. However, from the perspective of other interpersonal trust accounts they can be criticized as capable of ensuring only reliability, but not trustworthiness. Although, it is not so obvious but yet to be seen if in the institutional context reliability should be distinguished from trustworthiness.

Hardin's proposal of encapsulated interest view does not foresee trust in asymmetric power relations. This can serve as an argument to refute model of trust that does not offer a normative framework to incentivize and create places of trustworthy institutions. However, as we have indicated, the encapsulated interest view is just one dimension of internal motivations for trustworthiness, together with external and mixed inducements.

In order to further examine potential institutional mechanisms for ensuring trustworthiness as introduced by Hardin (2002) we take as a starting point his examination of trust in the context of institutional setting where he claims that "creating institutions that help secure trustworthiness thus help to support or induce trust" (Hardin, 2002, p. 30). Here he gives a particular attention to discussing cases of mistakenly speaking of trust while the issue is in fact about trustworthiness. That leads us to reframe the question of trust into the one of trustworthiness and to propose a new institutional model accordingly.

The suggestion of abandoning trust in large institutional settings where it would be replaced with different mechanisms (Hardin, 2002, p. 199-200) such as cooperation without trust (Cook et al., 2005) reflects two interconnected perspectives. On one side it fails to recognize the nuances of different modes of trust, not acknowledging the emerging trend of the strong thin trust mode of trust. This mode of strong thin trust, as we have shown, is present mostly in institutions that hold high values for us, be it our health, jobs, or important decisions about our lives. Apart from power asymmetry that can be present, the second crucial aspect is that it is often hard to have extensive evidence that would enable us to place our trust wisely. This constellation can therefore easily lead to losing trust in institutions, incentivizing exploration of new models of trust. One of the recent means of enabling trusting relations can be illustrated through digital technologies such as blockchain through decentralized structure that might prove to be better suited to enable trusting relations by encoding the model of trustworthiness in the software.

2.5 Conclusion

In this chapter, we addressed questions on trust in interpersonal relations, in institutions and digital platforms. More specifically, we were interested should we trust in digital platforms when the stakes are high instead of institutions? Is trust emotional or can it be based on evidence? Could institutions and algorithms be trustworthy?

In order to tackle those questions, we have conducted a critical analysis of different interpersonal and institutional concepts of trust and trustworthiness. First, we identified common features of concepts of trust such as three-part relation, distinction from the concept of reliance and specific property of a trust that as such cannot be willed. Then we focused on the particularities of main concepts of trust that diverge according to the motivational causes from goodwill account of trust, trust as affective attitude, encapsulated interest view and trust as social norms.

Based on our critical analysis of disparate approaches and concepts of trust we argue that in an institutional setting concept of trustworthiness should have a priority instead of he concept of trust. Placing trust blindly or misplacing trust in the untrustworthy can result in disappointment, misuse or even exploitations. Therefore, instead of emphasizing trusting relations, trusting the trustworthy should have a priority especially in the institutional setting. Based on the critical analysis of the concepts of trust we proposed three hypotheses that were elaborate in three separate lines of arguments.

In the first section based on the analysis of the goodwill account of trust (Baier, 1986; Jones, 1996) that is mostly used in interpersonal relations we argued that the concept is too narrow to be applied to the institutional context. Also, the very definition of goodwill account is missing clarity as the meaning of goodwill is very vague. But the main finding of our examination of the concept pointed to the relevance of the strong thin mode of trust that has not been recognized, although it plays a crucial role when applying the concepts of trust to institutions.

Current approaches and definitions of trust indicate that they should be understood within different modes of trust. In this new framework, it becomes obvious that the strong thick mode of trust is mostly present in interpersonal relations. The interpersonal goodwill or affective attitude concepts of trust are capturing very well that strong thin mode of trust.

But both concepts become too narrow if they are to be applied to trustful and distrustful relations between publics and institutions. Secondly, the weak trust of cooperative relationships could also be present in institutions where stakes are not so high. But most importantly, what should not be overlooked is that in current complex institutional settings other modes of strong thin trust are in place.

These modes of trust are present when the stakes are high having a big impact on our lives. But at the same time, we do not have an adequate understanding of the problem in complex institutional settings, as they are not grounded primarily on face-to-face relations where we can be more confident in deciding whom to trust. This increased presence of modes of strong thin trust (Hosking, 2014) leads us to conclude that a new prospect of ensuring trustworthiness of institutions is required.

We, therefore, analyzed different modes of trust introduced by Hosking (2014); namely strong thick trust, strong thin trust, weak thick trust, and weak thin trust. The distinction between the modes of trust in Hosking's (2014) terminology can be understood as follows. The strong mode of trust is present when the risks at stake are high and values that we are committing are high. Dealing with this strong mode of trust in interpersonal, face to face contact give us more confidence that we will place our trust well due to extensive evidence that we can acquire in a frequent and direct face to face contact. This strong thick trust can, therefore, more easily rely on the goodwill of the one we trust because we have more resources and evidence on competences and goodwill of the trustee.

However, if we try to directly apply the goodwill account in the context of larger institutions this strong thick mode of trust becomes impossible to successfully apply to an institutional setting. Arguing goodwill account to be too narrow to be applied to institutions, however, does not imply that trust is impossible in the institutional setting. Such a claim would disregard the important differences in the modes of trust activated in different settings and predominance of strong thin trust in institutions. Although the dominant concept of interpersonal trust cannot be applied to institutions, it does not imply that trust should be substituted by other mechanisms. Instead, it might rather require a new model of trustworthy institutions that should be applicable in the new context of pervasive strong thin trust.

In the second section based on the analysis of the concept of trust as an affective attitude (Jones, 1996, 2013) we examined more closely emotional aspects of trusting relations that can

lead to beliefs resistant to evidence. Based on the critical analysis of this concept we proposed a balanced approach between emotional and cognitive aspects of trust instead.

On one side Jones introduces the cognitive element of trust (Jones, 1996) that has to be grounded on the evidence about trustee's competence, goodwill, and responsiveness. It is contained in the expectation that the one who is trusted will be favorably moved by the thought that someone is counting on them. However, Jones does not assign priority to that cognitive part in the concept of trust as an affective attitude, where rather an affective character of trust takes a predominant role.

Based on the introduced case study in the context of science we rejected the adequacy of the central role of affective attitude within the model of trustworthy institutions, although emphasizing at the same time the relevance of the emotional aspects of trust that play significant a role next to the cognitive judgement. Importance of the affective attitude was evident since the trust in scientific claims proved not to entirely be dependent on the quality of evidence but often influenced by our pre-existing trust in the source who conveys the message as suggested by the work in experimental psychology (Kahneman, 2013) or social psychology (Kahan, 2013).

However, giving priority to emotional responses in the institutional context would hinder wisely placing trust and could have misleading effects that can result with harmful consequences for the public in the institutional context. Therefore we argue that in the model of trustworthy institutions both cognitive aspects of trust and emotional traits should be in interplay.

In the final section, we analyze the concept of trust as an encapsulated interest according to which trustworthiness is motivated by encapsulating the interest of the trustor in order to maintain the relationship (Hardin, 2002, 2006). Instead of interest as a main motivation for the trustworthiness, other motivations might be explored such as responsibility or responsiveness in the context of the trustworthy institutions.

Focusing on the proposal of abandoning trust in institutions based on the understanding that the cooperation is possible without trust, points to two interconnected points. First is failing to take into account new emerging modes of trust such as strong thin mode of trust that is present in complex institutions where the stakes are high but extensive evidence that would enable us to place trust wisely is lacking. The second implication is having an influence on easily losing trust in institutions and exploring new modes of trust provided by decentralized digital technologies that might enable trustworthiness by design. Based on acknowledging these prevalent modes of

strong thin trust and new technological modes of ensuring trustworthiness in the next chapter our aim is to formulate the model of trustworthiness that could accommodate concerns and inadequacies of the presented concepts of trusts and be applicable both to institutional as well as in the context of digital platforms.

CHAPTER 3: PRIORITY OF THE CONCEPT OF TRUSTWORTHINESS

Based on the results of the analysis of different concepts of trust that suggests prevalent strong thin trust in institutions, we attempt to articulate a new trustworthy by design model that should be applicable in this new institutional as well as in the context of digital technologies. Instead of exploring ways of restoring proclaimed decrease of trust in institutions, we take a different stance reframing the question in terms of trustworthiness. The conceptual framework of trustworthiness we base on the work of Onora O'Neill (2013) that suggests the possibility of placing trust well only if we trust the trustworthy. Therefore, priority will be given to the concept of trustworthiness over trust because it has distinguished properties that could be reflected upon when placing trust wisely.

Introducing the concept of trustworthiness enables us to apply a more selective approach to trust by suggesting criteria that should be met when singling out relevant properties that would make institutions or digital platforms trustworthy. In one aspect we part from O'Neill's line of argument by recognizing the role that values, emotions, and affective attitude can play in the context of trust and trustworthiness, although we do not suggest giving it a priority as it could result in placing our trust blindly. Our focus on the priority of trustworthiness first clarifies intentional characteristics of the concept and normative basis that is in interplay with empirical aspects of research.

We start from the perspective of trustworthiness, unlike most of the other concepts focused on trust, which might be relating to the concept of trustworthiness only indirectly. Few worthy exemptions from that rule, however, is Jones's (2012) reverse approach that places trustworthiness in focus, and further work on trustworthiness that takes into account personal and institutional articulation of trustworthiness (Potter, 2002; Hardin, 2002; O'Neill, 2002a, 2002b, 2013, 2014). Relying on this valuable work, we introduce several distinctive features to the current conceptual framework in search for more effective and adequate application in the contemporary context of institutional setting in the context of scientific institutions and algorithmic systems.

It should also be noted that giving priority to trustworthiness in this context of institutions, however, does not exclude the possibility of trusting without relating to the judgment of trustworthiness in specific interpersonal cases. The intention is not to replace other modes and

concepts of trust that are indispensable in other interpersonal contexts such as therapeutic trust or various manifestations of strong thick trust. Here, we agree with Hawley that *"*there can be many reasons to trust in the absence of trustworthiness, including therapeutic trust aimed at cultivating trustworthiness, psychological self-protection, efficiency in low-risk situations, and so on. Nevertheless, even in such cases, trusting involves behaving as if the recipient were trustworthy" (Hawley, 2017, p. 6). When it comes to the institutional context, however, those particular cases lose their relevance and mechanisms based on the concept of trustworthiness are necessary.

Before formulating the main institutional design properties of our model, together with the necessary mechanisms that should enable both signaling and responsiveness or uptake, we will first start with more general features of the concept of trustworthiness, as a foundation for claiming priority of trustworthiness. First, we will tackle the features of intentionality and normative approach and then properties of institutional design and mechanisms that should enable judging trustworthiness.

3.1 Trustworthy institutions model, Intentionality and Normativity

3.1.1 Intentionality

If our intention is to wisely place trust in trustworthy institutions, we have to be able to judge where to place trust based on the assessment of the properties that make institutions trustworthy. For this objective affective attitudinal character of the concept of trust that we have previously analyzed (Baier, 1986; Jones, 1996) seem to be problematic. It gives precedence to affective aspect that might also relate to belief resistant to evidence, thus undermining the prospect of intentionally placing trust. In order to support our claim of a close relation between intentionality and giving priority to trustworthiness when placing trust in institutions, we will contrast two different approaches. First, is the attitudinal approach introduced by Baier (1986) and Jones (1996) and second, O'Neill's (2013) proposal on intentionality that we adopt.

As we have seen, both Baier's and Jones's accounts understand trust as an attitude that cannot be adopted at will and is grounded on competence and goodwill. But as an attitude, trust can also be a belief resistant to evidence (Jones, 1996) that as such loses its ground to be intentionality based on rational assessment of trustworthiness. Nevertheless, it does not trump the importance of taking affective attitude of trust into consideration. As we have illustrated in the cases of mistrust in science regardless of the extent of the evidence provided.

In this example the reason for failed trust in scientific claims and scientific institutions was mostly due to failing to take up publics values regarding specific topic of their concern. This affective aspect will also prove to have a significant role in our formulation of the model of trustworthy institutions. However, by taking the affective and attitudinal aspect into account, we would still want to make sure that the conditions for wisely placing trust based on judging specific properties of trustworthiness of institution would have predominance.

At this point, we depart from attitudinal approach and contrast it with O'Neill's (2013) approach singling out its three distinctive features that we find crucial in addressing the possibility of judging trustworthiness in the institutional setting. Those are namely intentionality, interdependence between trust and trustworthiness and normative approach to trust. In regards to intentionality O'Neill's approach "rejects an "attitudinal" account of trust, since she takes trust

something that is “intentionally” placed in the trusted, where the placing may or may not be intelligently done” (Baier, 2013, p. 181).

Intentionally placing trust undoubtedly cannot ensure intelligently placing trust, but it surely enables the prospect of intelligently or well-placed trust. It also implies that we can judge specific properties or characteristics of institutional design through available mechanisms, be in the form of openness, accountability or other. O’Neill (2013) also emphasizes that “for practical purposes we are interested not simply in trusting attitudes, but in placing trust well - and any account of well-placed trust depends on an account of trustworthiness” (p. 237). This prospect of judging trustworthiness in order to intentionally place trust thus implies the second aspect of her approach of interdependence between intentionality and prospect of trustworthiness.

However, judging where to place trust can surely be a greater challenge in a complex institutional setting than when trusting another person. O’Neill (2013) acknowledges that “placing trust in another for some particular purpose we typically need to make a judgement of the other person’s competence and honesty, as well as their reliability, in the relevant matter” (p. 238). Making judgement of another person’s competence, reliability and possible other properties that are relevant for the specific domain in which we trust the person, is undoubtedly easier than applying the same approach to complex institutions. This becomes even more evident when it comes to strong thin trust, as we have previously described when the risks of trusting are high and the values we are committing great, but we do not have the capacity to examine all relevant aspects in the complex institutional setting we are dealing with.

We will adopt O’Neill’s framework on trust and trustworthiness, as it engages in interplay several elements that will be necessary for the formulation of the priority of trustworthiness in institutions. First of them is intentionality of placing trust that acknowledges the importance of placing trust well, it is intrinsically connected to the concept of trustworthiness and has important implications for the prospect of trustworthy institutional design.

3.1.2 Normativity

Normative aspect of trust is frequently left out from interpersonal and empirical approaches which mostly take empirical perspective in determining if someone trusts and how

much they trust. Although most of the research on trust has been done on empirical grounds, work of Onora O'Neill stands out when it comes to normative research, that should be equally important. Her stance on some empirical research and the relevance she gives to guiding principles formulates a starting point for our take on the normative aspect of research on trust. Furthermore, her suggestion of bringing the topic of trustworthiness not only to personal but also to social and political level (O'Neill, 2002b) enables the prospect of modelling trustworthy institutions. O'Neill (2013) clearly articulates her critical stance on some empirical research:

> "In many contexts, particularly in commercial and political life, where reputation is critical to success, information about others' levels of trust can be both useful and economically valuable. Some excerpts from such information are widely reported in the media in developed societies; others remain closely guarded in commercial or political confidence. However, empirical work on attitudes of trust does not offer an adequate framework for addressing normative, including ethical, questions about intelligent placing and refusing trust. In working out whom to trust for what purposes it is of little help to know whom others trust, or how much they trust them" (p. 237).

The results of measuring public trust the way it is mostly performed, although potentially beneficial for the reputation of public officials or marketing purposes, in terms of placing trust well and trusting more the trustworthy, would require another approach at the level of normative inquiry.

O'Neill (2002a) elaborates on achieving normative approach by translating ethical principles and high standards into legislation, regulation, policies and ensuring compliance. Although abstract ethical principles underdetermine action and implementation in policies as they are to be implemented along with other requirements (be it technical, scientific, legal, social, clinical, financial, political, etc.). O'Neill argues that principles have an important role in guiding our practical judgment in the implementation of policies. Principles cannot be "life algorithms", she continues, and they cannot substitute our practical judgment in action but are valuable guidance in design. Exactly as principles cannot be life algorithms, we can also not expect that from regulations and various mechanisms that should ensure compliance in the institutional context. Because not everything can be regulated, we do need guidelines and thus their implementation cannot rule out the necessity of trusting relations in institutions.

The very act of placing trust implies we can never have absolute certainty that we won't be betrayed, but that does not contradict the intentionality of placing trust wisely and implementing mechanisms in institutional design that would enable us to do so. As O'Neill (2002a) continues

> "If blanket scepticism is not a feasible basis for life we must place trust selectively and with discrimination even when we lack any guarantee that agents or institutions of any specific sort are unfailingly trustworthy. The possibility of being mistaken, deceived and even betrayed cannot be written out of life. It is therefore important to find at least approximate ways of distinguishing between well-placed trust and misplaced trust" (p. 122-123).

Bringing this challenge to the normative level of intentionally placing trust brings us to the next step of articulating the model of trustworthy institutions with the prospect of enabling us to judge how to place our trust well.

O'Neill's normative approach deals with ethical questions and practical purposes of trust, Its relevance is also in bringing the question of judgment on trust not only to the individual level, but also to the social and political level by emphasizing that "we need social and political institutions that allow us to judge where to place trust" (O'Neill, 2002b, p. VII). This requirement will be elaborated and taken on board in the later section addressing the challenge of trustworthy scientific institutions.

3.1.3 Trustworthy institutions model

We propose a model of trustworthy institutions as a three-part relation in which B is trusted by A for specific thing X. Where we hold B to be an institution holding necessary properties of the institutional design, namely competence, reliability, responsibility, and responsiveness regarding the specific thing X. Furthermore, a revised form of mechanisms of openness and accountability should enable trustworthy institutions not only to signal its trustworthiness to trustor, but also to be responsive in a way to also include uptake of trustees potential concerns and values regarding the specific thing in question. Our model of trustworthiness is thus extended so that trustworthy institutions not only signal to trustor their

trustworthiness but also include integral part of responsiveness in a form of uptake of concerns and values that A holds regarding specific thing X.

This extended model of trustworthiness is not only signaling - as signaling can in some cases fail - but also by responsiveness in the form of uptake of the concerns and values of the trustor, provides a missing link to the successful signaling and rich trustworthiness of institutions. Namely, as Jones (2013) rightly observes, rich trustworthiness can fail due to no uptake on behalf of the potential trustor, be it audience or the publics because of affective aspect of trust. However, we claim that this affective aspect of trust is not present only on behalf of trustor in form of their values, concerns, and biases, but also on behalf of trustee be it scientific institutions, policy makers, etc. because they are also subject to values and biases. In order to enable trustworthiness of institutions this dual side of potential affective elements should be acknowledged and responsiveness and uptake of potential concerns and values of trustor should be included.

The model can be applied to scientific institutions, where distrust in science may not be explained as insufficient understanding or knowledge deficit about the subject on behalf of publics (Goldenberg, 2016). Distrust and hesitancy are rather motivated by rejecting values that are behind research or algorithms and concerns about specific side effects that are not taken on board by scientific institutions as a part of research agenda. Therefore, signaling trustworthiness on behalf of scientific institutions in terms of competence, expertise, and commitment to improving health, for example, does not necessarily ensure uptake on behalf of potential trustor and publics may still distrust scientific institutions.

Applying the model of trustworthy institutions in this specific case should include both signaling and uptake of publics' concerns. Signaling trustworthiness on behalf of medical or scientific institutions involved in the research would have to be responsive and take up concerns of the public and deal with them if they want to be trustworthy and trusted. It might mean that, based on public concerns, research agenda will have to be changed to focus on the aspects of side effects that are most worrisome for the public's and provide answers or insights that are most problematic from the point of view of the public's. It may include also some other relevant steps based on the uptake of public's concerns. This additional point of responsiveness and uptake of trustors concerns as an integral part of signaling trustworthiness of the institutions will be further explained within the properties of the proposed institutional design.

3.2 Properties of institutional design, Responsibility and Responsiveness

As it becomes clear from our previous analysis, models of trust that we have analyzed cannot simply be applied to complex institutions. However, that necessarily does not lead to a conclusion that the prospect of trust should be abandoned. The concept of trust should rather be further extended in order to accommodate appropriate mechanisms and institutional design that would enable and incentivize trustworthiness. In our model of trustworthy institutions, we are proposing properties of institutional design in form of competence, responsibility, and responsiveness regarding the specific thing X, that would enhance the prospect of institutional trustworthiness.

Judging those properties in institutional design could be paralleled to judging other people's character, competence, and reliability - in the specific domain of our interaction - enabling us to place trust wisely. In the final instance, our goal shouldn't be to incentivize more trust - as that can lead to further misuse of our trust by institutions -but instead placing trust in the trustworthy. That is also the basis of O'Neill's (2002a, 2002b, 2013) understanding of the account on trustworthiness according to which:

> "In every domain of life we aim and hope to achieve not simply more trust, but specifically more trust in the trustworthy, and (if we can) less trust in the untrustworthy. What we care about is *well-placed* trust; what we fear is *misplaced* trust. So there is little prospect of saying much about the practical problems of trusting if one is unwilling to say anything about trustworthiness. Yet it is quite common for current discussions of trust to say remarkably little either about trustworthiness, or about the demanding task of distinguishing trustworthiness from untrustworthiness" (O'Neill, 2013, p. 237).

However, as we have already emphasized, placing trust in the trustworthy becomes even more challenging when applied to institutions.

Within institutions, it is not only harder to judge where to place trust, but we might even not be in a position to choose where to place trust, because institutions often cannot be substituted. It is much easier to choose to place trust in one or another doctor but when it comes to placing trust in governments, for instance, we might have fewer choices. As Weinstock (2013) argues, institutions have power over us and can hardly be substituted, so we have no choice but to "interact" with them in our interests. We depend on the knowledge of health institutions for our

health, for example, they hold our vital interest sometimes even in matters of life or death. So it is not always possible to employ "exit strategy" and to choose not to trust. The risk of mistreatment and exploitation in this context between institutions and publics is, therefore, greater than in interpersonal context and requires a different approach. Further challenges regarding trust are moving into digital sphere, trust in institutions possibly being displaced by trust in digital platforms where decisions are often based on algorithms, thus posing new challenges for the trustworthiness of the institutions.

The risks and possible negative consequences that are associated with trust in the context of institutions might even be greater than in interpersonal trust making a question about incentivizing trustworthiness even more relevant. Incentives for trustworthiness in institutions might be similar to the climate in a political and social setting that is conducive for trustworthiness or on contrary might have an impact on distrustful default position. Democratic societies would create different climates from the oppressive ones, thus influencing different default position on trust (Govier, 1997; Uslaner, 1999; Walker, 2006; Welch, 2013). Different climate can then either incentivize trustworthiness or induce default position of distrust.

Along the same lines, Jones (1996) does not think it is possible to give a general answer to the question on what would be rational default position *on trust*. Would it be the one of trust, distrust or neutrality? She rather perceives it as dependent and sensitive to four criteria climate, domain, consequences and the one that is agent specific. Among those criteria, she identifies climate as the decision on justified trust in some particular case that is dependent on the general features of the social climate. Some social climates that provide strong motives to be untrustworthy would require more evidence to justify our trust, as in the example of the Chinese Cultural Revolution, where defaults position would be one of distrust. On the other hand, we would need less evidence to justify our trust within the environment that holds little incentives to be untrustworthy.

We will apply the elaborated notion of relevance of the sensitivity to the climate, to the level of institutional design in order to emphasize why specific features and properties of institutional design might have an important role as incentives for trustworthiness in institutions. We adopt the criteria of competence used across various concepts of trust together with the commitment as necessary ingredients of any prospect of trustworthiness. But we further extend

the formulation of trustworthiness in institutions to integrate our proposal of properties of responsibility and responsiveness as essential ingredients of institutional trustworthiness.

3.2.1 Responsibility

As we have argued, the concept of trust and trustworthiness are to be formulated as normative expectations that, in O'Neill's (2002a) words, could not serve as "life algorithms" but could only be guidelines. In order to be able to place trust well, we would have to assume that the trustee should be competent, responsible and responsive in the domain where we place our trust. We take responsibility to be one of the normative requirements that we expect from the trustee, the criteria rarely taken on board in the conceptualization of trust, with few exceptions (Walker, 2006; Ruokonen, 2013).

It would not be disputed that in order to be trustworthy one has to have competences in the specific field of expertise related to the domain in question. It should not be ignored however that competences do not have to apply only to expertise, but could also be related to moral competences such as responsibility for the actions in the domain of our trustworthiness. Each person should be responsible for their actions and their consequences. When it comes to institutions, be it financial, medical or scientific institutions, that requirement becomes even more important due to the impact on people's lives.

If institutions are incentivized by self-interest, as we have witnessed on numerous occasions, consequences on the publics can be immense, often resulting in losing public trust. In order to enable public trust, they should rather be incentivized to act in a responsible way. If seen as collective agents (List and Pettit, 2011; Weinstock, 2013), institutions could also be held responsible and their trustworthiness judged by that criteria. Weinstock relates responsibility to trust within the institutions arguing, "at least some institutions are agents, and thus, the kinds of entities to which it is appropriate to ascribe responsibility, and to evince attitudes such as trust" (Weinstock, 2013, p. 207).

Ruokonen (2013) also designates responsibility rather than goodwill as a motivational aspect of trust. She argues "goodwill does not explain the difference between reactions to breaches of trust and reliance" (Ruokonen, 2013, p. 8). Although she is mainly focused on the

personal account of trust, the responsible agent as she defines it can also be easily applied to institutions if we understand them as collective agents. Account of responsibility that Ruokonen suggests thus consists of the same basic understanding of responsibility for the consequences of our action as rational and reflective beings. She assigns two aspects of responsibility to agents "a responsible agent has the capacities of understanding, reasoning, and control of conduct. Secondly, an agent can be (retrospectively) held responsible for at least some events or outcomes which can be ascribed to one" (Ruokonen, 2013, p. 9). If all the aspects that she holds relevant to a personal agent, namely autonomy, rationality and reflectivity can be applied to the collective agent, then responsibility can be an integral part of the trustworthy institutional design.

Furthermore, it becomes clear that in the context of distributed complex systems that can apply to institutions or companies, the existing ethical frameworks for individual responsibility are no longer adequate. For those purposes, responsibility that applies to the distributed agency (Floridi 2013; Floridi and Taddeo, 2016) is better suited to hold "all agents of a distributed system, such as a company, responsible (Taddeo and Floridi, 2018, p. 751). As illustrated in the previous section it plays important role in the algorithmic decision making and AI where the responsibility should be shared between diverse actors involved, from designers, software engineers, and regulators to users. In the next section, we will further illustrate it by elaborating on Douglas's (Douglas, 2009) proposal of moral responsibility for individual scientists and extending it to institutional level in the context of scientific institutions.

3.2.2 Responsiveness

It should be emphasized that the concept of trustworthiness of institutions that we formulate from the perspective of the trustee, should not be taken as one-sided. The request for trustworthiness should be seen as dual-sided, both on the part of the trustee as on the part of the trustor. In the context of science and scientific institutions it would mean that public trust in science and scientific institutions, requires also trust in publics. Or in the context of algorithms that public trust in institutions and companies that employ algorithms also requires reciprocate trust in publics.

In the context of scientific institutions that would presuppose introduction of novel forms of co-production, from framing the question to agenda setting. Sloman and Fernbach (2017), along the same lines, imply that science holds just one part of the knowledge. Therefore, they suggest reaching out to other types of expertise and knowledge present in the community and local expertise. This diversified field of expertise should not be lost in excluding some voices from the side of the publics. Public engagement, along those lines, should also be reframed in order to stop "continuing failure of scientific and policy institutions to place their own science-policy institutional culture into the frame of dialogue, as possible contributory cause of the public mistrust problem" (Wynne, 2006).

This notion of co-production practices from the context of scientific institutions can serve as an illustration of enabling two-sided trust between publics and institutions, between trustee and trustor. A base of this two-sidedness is responsiveness, which we suggest is next cornerstone of trustworthy institutions design. We understand this term as responding to the concerns of public and society.

Being responsive in a way of taking into account values and concerns of the publics differs though from Jones's understanding of responsiveness. Jones (1996) understands the term as being "directly and favourably moved" by the fact that someone is counting on us so that the trustworthy is acting responsively because he has been counted on, although she admits there might be other motivations. This motivational reason cannot be applicable in institutional trustworthiness account as the motivation cannot be expected to come from the fact that the trustee is moved by the fact that someone is counting on them. More likely is to be motivated by the commitment to the responsible conduct towards the trustor and not by the simple fact that someone is counting on us.

Furthermore, responsiveness should always be closely related to a specific domain in question, as it could not be expected of institutions or individuals to be responsive to all the expectations or concerns that can be expressed. It should also follow the formula where B is trustworthy to A in specific thing X if he is also responsive to concerns and attitudes that A might have in specific thing X. As we have noted the formula should also be applicable in reverse way where A is trustworthy. Responsiveness to publics concerns should be at the very basis of the trustworthy design from the very beginning of the process thus ensuring the trustworthiness by design.

3.3 Mechanisms of institutional design, Accountability and Openness

After identifying properties of institutional design that we propose as a means of incentivizing trustworthiness that should be judged in order to place our trust wisely, we still face the challenge of the practical attainability of actually being able to judge trustworthiness in the context of institutions. Therefore, in this section we suggest that institutions should signal their trustworthiness similar to the signaling in the rich trustworthiness concept introduced by Jones (2013). As it became clear in differentiating thick strong and thin strong trust, the amount of evidence and information that we have or could obtain by ourselves is dramatically lower when faced with the challenge of judging the properties of trustworthiness of complex institutions than in interpersonal relations. Acquiring by ourselves the evidence on the competence of institutions or their responsibility and potential responsiveness often extends our capacities as individuals. Just as we cannot gain all the necessary knowledge by ourselves, but we rely on the expertise of others. We do not gain the knowledge that earth is round by ourselves but we rely on the expertise of others (Hardwig, 1991).

When institutions fail in their trustworthiness it can often be too late and damages to trust can be hard to repair. Hosking draws the analogy of establishing trust and confidence to coconut tree "as an Indian policymaker has commented, 'Confidence grows at the rate a coconut tree grows, and it falls at the rate a coconut falls" (Hosking, 2014, p. 2, chapter 3). Trust can very easily be destroyed but not as easily repaired. It becomes even more evident in the event of the disaster such as Fukushima nuclear disaster. After the reactions from the government and scientists that followed after Fukushima disaster, the impact on trust in governmental and scientific institutions were severely damaged:

> "One displaced university lecturer commented a year later, 'Since 11 March people haven't trusted scientists who receive funding from the government. They trust people who act without government funding and who work together with them.' A nation which traditionally trusted government implicitly refocused its trust on independent professional associations and voluntary groups" (Hosking; 2014, p. 3, chapter 3).

Taking into account complex challenge of judging the trustworthiness of complex institutions and also the consequences of breaches of trust in institutions, in order to enable judging where to place trust, we need institutions to signal their trustworthiness. In this vein,

Jones introduces the term of rich trustworthiness where trustee identifies himself as the one who can be trusted. In an institutional setting, it might seem that signaling might often be enabled through mechanisms such as accountability and transparency.

However, implementation of the mechanisms of accountability and transparency so far did not completely fulfill their purpose. They often either create perverse incentives (O'Neill, 2014) or are put in place in order to replace trust (Cook et al., 2005). As O'Neill rightly observes "Trusting is not a matter of blind deference, but of placing - or refusing - trust with good judgement. So we need social and political institutions that allow us to judge where to place our trust. Yet some fashionable ways of trying to make institutions and professionals trustworthy undermine our abilities to place and refuse trust with discrimination"(O'Neill, 2002b, p. VII-VIII). O'Neill (2002b, 2014) provides further analysis and critique of both candidates that have been introduced in forms of accountability and transparency in order to monitor institutions and provide means of judging their trustworthiness.

In what follows we will first examine the context of signaling trustworthiness in institutions, second we will focus on identifying the flaws of the mechanisms of accountability and transparency. In the final instance, the analysis should enable us to identify the mechanisms that would enable publics to judge the trustworthiness of institutions in order to place more trust in the trustworthy and less trust in the untrustworthy institutions.

3.3.1 Signaling trustworthiness in institutions

In our proposal of signaling trustworthiness in the institution we take Jones's (2013) introduction of rich trustworthiness as a starting point that is defined in terms of signaling trustworthiness. Then we part from Jones's concept of rich trustworthiness in two main ways, first by introducing the notion of the intentionality of trustworthiness and notion of responsiveness where the concept would be applied simultaneously to both trustor and trustee.

By applying the concept of rich trustworthiness to institutional context our model of trustworthiness in institutions presumes more active form of signaling trustworthiness that as such has also the quality of intentionality. It is intentional through the properties of institutional design that would enable publics to be in a better position to place trust wisely. It is not so

obvious that the process of signaling would happen by itself and that institutions would identify themselves as trustworthy as could be implied from Jones's understanding of signaling. In an institutional setting it has to be intentionally enabled by putting in place mechanisms, such as openness and accountability that would enable it. In our model of trustworthiness, signaling does not only imply intentionality but also extends to responsiveness thus ensuring uptake of trustors concerns and values regarding the domain in question.

The second main way in which we depart from Jones proposal is in the context when signaling can fail. Based on the analysis of failed signaling we suggest that both emotional and cognitive character of trust have to be acknowledged not only on behalf of trustor but also on behalf of the trustee. The trustor here can be understood as publics or in the context of Jones proposal as audience, while the trustee can be institution, company or scientific institution where the emotional aspect comes in the form of values or potential biases.

In formulating the model of trustworthy institutions that includes uptake of values and concerns of trustor in a specific domain of trustworthiness, we adopt a revised version of Jones's (2012, 2013, 2017) account of rich trustworthiness. Karen Jones distinguishes basic from rich trustworthiness, in a way that the former also signals its trustworthiness. The act of signaling to others has a purpose of indicating that one can be trusted and relied on. Jones specifies that signaling can be done in various ways by deeds, words, certificates or institutional roles. Precisely, Jones understands signaling as follows:

> "Signaling is communicative and takes place against a vast social background including norms and shared understanding of what can be expected of whom. Some of the background norms are moral, others professional; some are localized, others broadly based… Because we live in a world in which how we present ourselves and who we are taken to be carries with it social meanings, we are inevitably signaling, rightly or wrongly, who can rely on us for what. … Signaling can be accomplished in many ways, including but not limited to: appearance, comportment, words, glances, deeds, certificates, titles, institutional roles" (Jones, 2013, p. 190).

The first question that comes to mind when applying the term signaling could possibly be lack of a clear distinction between signaling and communicative act as suggested by O'Neill (2013). However, Jones definition suggests that the main distinction could be found in the fact that

signaling does not only refers to words but also to other acts of showing trustworthiness by actually doing something and by deeds.

Jones thus introduces the form of rich trustworthiness where signaling enables the trustee to identify himself as trustworthy thus extending trustworthiness to its richer version where

> "B is richly trustworthy with respect to A just in case: (i) B is willing and able reliably to signal to A those domains in which B is competent and will take the fact that A is counting on her, were A to do so, to be a compelling reason for acting as counted on and (ii) There is a nontrivial number of relatively central domains in which B will be responsive to the fact of A's dependency in the manner specified in (i)." (Jones, 2017, p. 8)

Where we part from the account of trustworthiness as elaborated by Jones, is its possible implementation in the institutional context that holds the aspect of intentionality. The concept of rich trustworthiness can be a further active step towards ensuring trustworthy relations, although it could be applied not only in terms of interaction between individual actors but might adequately be used in the institutional context by applying mechanisms such as openness and accountability. Furthermore, in institutional setting signaling also has to be intentional that requires reexamining Jones's claim that "In order to signal, we do nothing at all" (Jones, 2012, p. 76). Signaling cannot happen by doing nothing at all, because it is enabled by the mechanisms intentionally put in place such as transparency or accountability. In both cases, it requires mechanisms and infrastructures of providing open data and information or procedures of how to hold institutions to account.

The second aspect of Jones's concept of rich trustworthiness that is of particular relevance in the context of institutions is that signaling can fail or be exploited. We will suggest that both of those critical points could be minimized within the model of trustworthy institutions. Namely, as we have previously elaborated, request for trustworthiness should be mutually reinforcing for both trustee and trustor. It should also engage in interplay both cognitive and affective aspects thus enabling the model of trustworthiness to recognize affective aspects of trust related to both trustor and trustee and enabling uptake of this emotional aspect, as for example trustor's concerns and values. We will first focus on Jones's arguments related to failed signaling and then elaborate suggestions of minimizing them in the context of trustworthy institutions.

Jones proposal of the concept of rich trustworthiness can be one of the possible solutions in solving misplaced mistrust that can be described as "...something missing from the trustworthy person or institution that is unable or unwilling reliably to signal to would-be trusters that they can be counted on" (Jones, 2013, p. 187). But it's success also depends on overcoming tangible obstacles to the uptake of rich trustworthiness. Jones recognizes that rich trustworthiness fails in communication when there is no uptake from the audience due to the character of trust that seems not only to be judgment and choice but is also unreflective and emotional.

Jones's points out that the character of trust is unreflective and includes emotional aspects such as possibly prejudices, cognitive biases, social experience and so on and can not be seen only as judgement and choice that we make about trustworthiness. But apart from assigning the emotional character of trust to trustor (be it publics or the audience), the trustee who signals its trustworthiness (be it person or institution) should also not be seen only as acting in a reflective and rational way. Failing in the signaling of rich trustworthiness due to the emotional aspect of trust should not be understood as one-sided and attributed only to the unreflective character of trust on behalf of the audience or publics alone. This narrow way that distinguishes rational signaling of trustworthy person or institution on one side and emotional aspect of trust on the part of the audience, on the other hand, can hinder the prospect of uptake of trust.

In the context of science institutions failing to recognize this affective aspect of signaling trustworthiness as values that can be traced both on behalf of trustor as well as trustee can lead to a trap of the knowledge deficit hypothesis. Claiming that if only publics, or the audience would understand the signaling of trustworthiness, the trust would be achieved. Or as formulated by Wynne in terms of failed imagination "an apparent institutional lack of ability to imagine that public concerns may be based on reasonable questions that are not being recognized and addressed, rather than being rooted in ignorance and misunderstanding" (Wynne, 2006, p. 219).

This topic of public mistrust in science clearly illustrates that in signaling trustworthiness there is an underlying affective layer of values that can be observed both on the behalf of trustor as public and as well as the trustee as a scientific institution. The character of trust is partially unreflective on both sides due to underlying values that are intrinsic to both.

> "Bearing in mind the institutional reflexes of continual reinvention of what is effectively an alibi projected onto others, and given the ab initio judgement of public mistrust of 'ourselves and our own' as unjustified, hence somehow misinformed, one can understand

> how this persistent institutional projection and reinvention occurs. Since it appears to be so creatively resistant to simple empirical contradiction, it has to be seen as reflecting a deep institutional-cultural need rather than a deliberated deception. It has been cumulatively entrenched over decades and energized by profoundly emotive feelings and insecurities about power and authority, emotions whose denial by reference to reason only make it all the more alienating and incoherent" (Wynne, 2006, p. 216-217).

Since the character of trust both on behalf of trustor as on behalf of trustee has partially affective character rich trustworthiness would have to adopt its richer version to acknowledge that dual affective aspect on both sides. Apart from signaling, the responsive and co-creative aspect of being open to an uptake of the concerns and values of the trustor, be it audience or publics should not be neglected.

Once both emotional and cognitive character of trust on behalf of trustor and trustee are being acknowledged, a new challenge is how to address them. O'Neill (2013) proposes dealing with pathologies of trust and failed communication - or in Jones terminology failed signaling - through epistemic and ethical norms that are needed in order to clarify the steps for correcting prejudices that are inherent within emotional and affective aspects of trusting relations.

Within experimental psychology (Kahneman, 2013) considerate attention has been given to this emotional and unreflective aspect of trust. It provides valuable insights into our natural tendency that when trusting we rely on our intuitive judgment or System 1. Although that process comes easily and effortlessly it can also expose biases and can be false. Kahneman research shows how important it is to give a due attention to reasoning and to slow down when trusting, specifically when it comes to important decisions. To satisfy O'Neill's request of placing trust intelligently we have to first overcome our natural tendency to rely on intuitive System 1 heuristics that can fail us or even lead to exploitation. Instead, we should employ reasoning in determining properties of trustworthiness such as competence, honesty, and reliability, as O'Neill would propose.

Focusing on the intentional aspect of signaling trustworthiness, we come to a conclusion that signaling can also fail, that lead to the introduction of its richer version. In the next section, we will examine the second challenge when signaling can be exploited. Within the model of trustworthy institutions exploited signaling means that mechanisms of transparency and accountability can be exploited. So we will first need clear formulation of both of those two

mechanisms of accountability and transparency as a means of signaling trustworthiness in the institutional settings in order to review the cases of their exploitation.

3.3.2 Accountability

Judging trustworthiness of complex institutions might be too big of a challenge for the individual making it harder to place trust wisely and judge the properties of trustworthiness in institutions. Therefore, it might be proposed to replace trust with cooperation or other mechanisms such as accountability. Although the interpersonal account of trust cannot be applied to institutions it does not necessarily imply that trust and other accounts of trust have to be replaced by mechanisms such as accountability. Not all aspects of conduct in institutions could be prescribed in detail or micromanaged, but only be in form of guidelines. Since guidelines cannot be prescribed or offer guarantee of conduct, within institutional design trust is still required. To elaborate on that point we will examine the current mechanisms of accountability that should be in place to substitute trust. Our analysis will start with O'Neill's (2014) critique of the mechanisms of accountability that will then take place in the broader context of features of an institutional design compatible with the presumption of institutional trustworthiness.

The question of trust in complex institutions by individuals and publics has recently received greater attention, often leading to questions of increasing loss of trust in institutions. Applying the interpersonal modes of trust as a means of judging the trustworthiness of institutions, as our analysis has shown, cannot provide adequate tools, implying that other modes and mechanisms are required. Judging properties of trustworthiness in complex institutions might not only pose a serious challenge for individuals, but it might seem almost impossible.

In the context of scientific institutions, it is best illustrated by the challenge of trust in the expertise of experts and scientific institutions. O'Neill (2002a), in this regards, rightly questions how shall we expect individuals to judge claims of scientific experts who have highly specialized knowledge in science, medicine etc. It is surely not easy to make judgement about institutions that are becoming more and more complex. It has become much harder for ordinary individual or publics to judge properties of trustworthiness of institutions and particular actors in order to determine where to place trust and for what specific purpose.

Furthermore, the concept of trust that originates in interpersonal relations cannot be applicable in the context of complex institutions which might lead to the conclusion that trust is not possible in complex institutions. As an account that originates in the interpersonal, face-to-face relations it surely cannot be the basis of intelligently placing trust in the institutional setting. O'Neill (2014) joins us in acknowledging the inadequacy of such an interpersonal account in the context of institutions

> "It is rather obvious why this conception of trust looks questionable in complex institutional settings, where relationships are usually neither face-to-face nor one-to-one, and where affect and attitude are not prime considerations. … A better account of trust that sees it as an intelligent response to evidence of trustworthiness is needed …" (p. 3).

If the interpersonal account of trust is inadequate for this purpose, that doesn't exclude possible better prospects of different accounts of trust that should be more appropriate when dealing with present challenges.

Secondly, we will address the suggestion of replacing trust by the mechanism of accountability based on O'Neill's (2014) critique of the currently applied concept of accountability that is managerial and focused on micromanagement. O'Neill's criticisms of the mechanisms of accountability that in their current form of implementation do not contribute to the institutional design that would enable publics to place trust more wisely. O'Neill (2014) argues that because of the proclaimed crisis of trust we abandoned the traditional approach of compliance that was relying on the culture of trust and instead we have replaced it by more formalized structures of accountability. This new managerial approach of accountability is intended to replace trust because trust that relies on interpersonal relations and goodwill is not attainable in complex institutions.

O'Neill at this point, also rejects both accounts of trust as attitude and managerial approach to accountability proposing instead a more intelligent concept of trust and a more intelligent concept of accountability. Her proposed intelligent mechanism of accountability should enable informed and independent judgement on behalf of an expert that would be pursued without conflict of interest. Finally, the last aspect in our analysis of O'Neill's proposal will be extended in terms of institutional learning based on the failed previous actions.

The concept of accountability that should substitute trust in institutional setting O'Neill calls the managerial conception of accountability. This managerial approach sets targets as

standards for controlling the performance of institutions and individuals. As such it is a ticking box approach in measuring success, meeting targets and consequently sanctioning or rewarding. The argument for its use is that it is cheap, objective and transparent, although all of those elements could be disputed. O'Neill claims that avoiding expert judgement does not make it objective, measuring properties that are not directly related to what we are interested in does not make it cheap and the fact that indicators for different rankings are easily transferable for the purpose of transparency does not contribute to intelligible communication.

On contrary, it can often lead to oversimplification and create perverse incentives. For some actions, it is not easy to set targets so instead proxy indicators are used. They might be easily measured but also could be over-simplistic and misleading. Furthermore,

> "Even well-chosen performance indicators can create perverse incentives. If exam marks are taken as an indicator of school achievement, this will have unintended and ultimately perverse effects. Pupils will be guided into the areas where high marks are easier to get; exam performance will be stressed at the expense of other educational objectives; schools will find ways to 'game' the system" (O'Neill, 2014, p. 7).

Coming back to our question about possible ways of enabling publics to judge the trustworthiness of institutions, managerial forms of accountability does not seem to provide the necessary tools for informed judgement on the trustworthiness of institutions. O'Neill argues that "The basic reason why managerial forms of accountability are misdirected is that they often make it harder to judge whether claims or commitments are trustworthy, so do not support but undermine the intelligent placing and refusal of trust" (O'Neill, 2014, p. 11).

She proposes instead, different, intelligent structure of accountability that would support intelligently placing or refusing trust. Because her account of intelligent and well-placed trust requires discrimination, an intelligent model of accountability has to enable providing the evidence necessary for the informed judgement of the performance of institutions or individuals. Evidence could never be complete and no firm guarantee could be granted because if that is the case we would not have to place trust at all.

O'Neill's intelligent account of accountability has three distinctive features, namely, "the benchmarks for intelligent accountability are informed and independent judgement of performance, complemented by intelligible communication of those judgements." (O'Neill, 2014, p. 16). Intelligent accountability enables informed judgement by making it clear what is the

normative claim, what is required action, or in other words what are the obligations and duties of the institutions. It should be clear what ought to be done so that it can be compared with the actual performance, of what was actually done. This informed judgement on behalf of an expert is completely different from ticking the boxes approach for assessing performance indicators. Next to being informed, expert judgement on the adequacy of performance also has to be independent and not pursued if there is a conflict of interests. And finally, it must be intelligibly communicated performance of those who are held to account and the communication must be assessable by the audience and the publics. Such an account, however, extends far from mere transparency.

We can surely agree with O'Neill's account of intelligent structures of accountability, on the point that it should provide evidence on the performance of the institutions in order to judge their trustworthiness. In order to be able to make an informed judgement, we would surely want to know what are the institutional rules, what were the obligations and to be able to compare it to the evidence on performance and their fulfilments assessed by an independent expert who can communicate it in the intelligible and accessible way.

Mechanism of accountability that would function in this way would give us more means to judge fulfilment of required action according to the properties integral to trustworthy institutions. As O'Neill (2014) understands it "Required action may be of many sorts: we commonly distinguish, for example, between legal, institutional, instrumental, customary, and ethical requirements, and this list is not exhaustive" (p. 11). Although O'Neill does not elaborate in details all the possible forms of required actions, in the model of trustworthiness it would surely require competence, responsiveness, reliability, and responsibility in order to better judge the mechanism of accountability that O'Neill proposes.

The second challenge that needs to be addressed should consider not only trustworthiness of the institutions in terms of judging it, but also its properties of design that would enhance trustworthiness. Here we have in mind good institutional design or well-structured institutions as O'Neill calls it. In general terms, the institutional design would put in place mechanisms to incentivize competent, responsive and responsible conduct. Or in the way that O'Neill puts it, it would provide support for the fulfilment of obligations, enable growth of professional integrity, robust system for dealing with conflict of interests and remedies for failure.

The final condition of the institutional design that O'Neill introduces deals with the failure of required actions. This aspect might be of special relevance for the institutional design because it could serve as a basis for the institutional learning. As such it would also give more dynamic character to institutional trustworthiness instead of perceiving it as a fixed structure. This kind of dynamic trustworthy institutional character would make it more effective, resilient and open to possible input from the trustor. In this way, failure could also be understood also as possible criticism or expressing concerns. If potential failures or concerns would be taken up they might enable institutional learning and enrich institutional trustworthiness. Together with signaling institutional trustworthiness, being open to the possible uptake of the perspective of the trustor, enables also acknowledging their values as modelled in the trustworthy institutions model.

This prospect of learning and change in institutional design can be illustrated as a means of strengthening trustworthiness. "When problems arise, it is crucial that their source be adequately identified. Do they result from individual breaking rules, or on the contrary do they point to perverse incentives that are generated by flaws in institutional design? Ridding institutions of such design flaws is key to the ethical resiliency of institutions, and thus, to their trustworthiness" (Weinstock, 2013, p. 214-215).

Weinstock's example from the recent Canadian history further clarifies what can be the significance of institutional change for strengthening trustworthiness of institutions. As he elaborates, in 2003 Federal Auditor-General issued a report that addressed the corruption and misuse of public funds and recommended punishing those who were guilty of the corruption. The report further uncovered structural flaws, because the head of the public service was also holding a cabinet position in the government. The report urged the government to introduce constitutional changes in order to prevent politicization of public service, but the systems flows still have not been tackled. "The government's response, however, was to put in place exactly the kinds of micro-managing mechanisms meant to signal the government's seriousness in tackling corruption, but doing nothing to address the systemic flaws that impede the government's trustworthiness" (Weinstock, 2013, p. 215).

This indicative example can illustrate why mechanisms of accountability are better to be introduced at the meso- and macro-level rather than in the form of micromanagement. Micromanagement and ticking the box approach doesn't seem adequate for enhancing the trustworthiness of the institutions. O'Neill also criticizes the attempt to replace trust by

introducing managerial systems of accountability where performance should be measured in quantitative terms by ticking the box approach as it does not seem to solve the challenge of trustworthiness.

Other more general structures that would apply to meso- and macro-level institutional design seem to offer a more adequate approach to trustworthy institutional design, Weinstock suggests:

> "The agency of institutions is a product of institutional design: it results from the ways in which the many individuals who work within the institution in question interact in order to deliver the goods or services that the institution is designed to deliver, the rules that govern their interactions within the institution, and the various oversight mechanisms that enforce these rules. Accordingly, the trustworthiness of institutions should be assessed on the basis of meso- and macro-level design features of institutions, rather than primarily upon the apparent trustworthiness of particular individual human agents within these institutions" (Weinstock, 2013).

Focusing rather on more general features of the structural design on meso- and macro-level that would implement features incentivizing the trustworthiness and introduce changes in the institutional design as remedies for certain failure seems to offer a better prospect of tackling the challenge of trust and trustworthiness in the institutional context.

3.3.3 Openness

Another candidate mechanism for enhancing our trust and enabling us to judge the trustworthiness of institutions is transparency and openness. Although their positive aspects are undoubtedly necessary and beneficial, the possible distortive impact they might have could result in adverse consequences for the trustworthiness of the institutions. There are countless examples of the high relevance of openness and transparency, such as openness about values behind scientific claims that are relevant for the democratic accountability of scientific practices, or openness about normative claims of what ought to be done and what was actually done, or openness about constitution and monitoring mechanisms within institutions.

A beneficial aspect of openness and transparency of those and numerous other possible implementations in various social and institutional contexts are self-evident. However, uncritical acceptance of all possible accounts of transparency would surely be mistaken as transparency and openness sometimes can have a damaging impact on the prospect of trustworthiness of institutions. There are two main critical points regarding transparency that we want to focus on. First is the issue of what information and data are to be open and transparent and the second is related to the difference between transparency and communication.

Tackling the first question of what is to be openly accessible takes a new turn in the information age that has a huge influence on availability and usage of information, data and big data in so many parts of our lives. The most obvious one is surely media, but it also has a great impact in the opening up of the research data and research results, and finally to the new ways that decisions influencing our lives are made with the help of big data and algorithms.

The challenge for trust in an increasingly open and transparent environment of new information technology is surely not new, O'Neill already addresses it in her work on trust and transparency (O'Neill, 2002b, 2014). Its relevance, however, comes to its fruition more than ever today, in the abundance of available information that at the same time increases the amount of misinformation in the post-fact world. It makes it harder to tell fact from fiction and place trust reasonably, consequently also possibly contributing to the decline in trust.

O'Neill's critical analysis points out two main critical aspects of transparency that enables both deception and creates perverse incentives. Due to the urge to open up, information can be adjusted, reshaped, or repeated and it is often hard to actively check the source of information, even more so in the new social media environments. The possibility of fabricating information or providing contrary evidence she claims shows how it can be more a reason for, than an effective cure against deception. Secondly, O'Neill argues that transparency is also not always necessary for accountability, but can sometimes even damage it by creating perverse incentives. She gives an example where

> "Position papers may minimise or omit serious discussion of options that might get a bad press, depress share prices, or provide information that helps competitors. Minutes may be drafted to support public relations rather than to provide accurate working records. Even more damagingly, demands for ubiquitous transparency create incentives to do more

> outside meetings, whether in private conversations, as chairman's action, or in unminuted" (O'Neill, 2014, p. 14).

In particular, she is focusing on the distortive impact of choosing to publish some material and not the other. Also, the time spend on defense does not contribute to the intelligent form of accountability or enabling the trustworthiness of institutions. In the context of scientific institutions, we will conduct a more detailed analysis of those challenges in terms of opening up of research data and research results as well as possible practices of its misuse. The challenge tends to be how to address deception and active production of ignorance within the quest of transparency in trustworthy institutions.

The second point that we want to raise is related to the interrelation between transparency and communication. Instead of drowning publics in information that are made publicly available and transparent, we adopt O'Neill's suggestion of intelligible, accessible and assessable communication as a better prospect. Moreover, this communication when conducted *with* publics rather than *to* publics enables taking publics concerns into account. Thus within the model of trustworthy institutions, it enables not only signaling but also including uptake of publics concerns within the context of responsiveness.

Transparency of information and data might seem like a perfect means of enabling judging trustworthiness of institutions. Although the intention might be reasonable, the mere opening of the huge amount of data and information by putting them in the public sphere often can have opposite effect.. O'Neill (2002b) rightly perceives that drowning publics in information disables them to process it critically and claims (O'Neill, 2014) that transparency is never sufficient for accountability to the publics.

The mere availability of information does not ensure good communication with diverse publics

> "since transparency is only a matter of making material available, of disclosure, it does not mandate and often does not achieve good communication with specific audiences. Shovelling facts and figures onto websites is not usually a good way of communicating, except with fellow professionals who have the time and expertise to sift and use what is disclosed" (O'Neill, 2014, p. 15).

In order to ensure that a good communication with the publics takes place it has to be intelligently communicated and it should be accessible by the publics.

A further point that O'Neill introduces as successful communication *with* publics rather than *to* publics also envisages taking publics concerns into account and as such is a relevant aspect of the account of institutional trustworthiness. Firstly, good communication should happen with the publics and secondly, it has to take into account the specific concerns of the audience. The distinction between transparency and communication, therefore, could be understood insofar as the later holds the relation *with* the publics, while transparency only makes it available for presenting *to* publics. This interactive aspect of the communicative process with the publics can be further extended to the uptake of the concerns and values that publics might hold.

Those two aspects of communication with the publics and taking publics' concerns into account we take to be essential ingredients for the robust trustworthiness of the institutions. Insofar lies also the critic of insufficiency of transparency as a means of ensuring accountability and institutional trustworthiness. Drawing publics in information or the fact that information might be transparent or extensively provided on one matter but not on the other does not seem to enable institutional trustworthiness. Signaling trustworthiness within the communicative act and ensuring intelligible and accessible information also cannot be sufficient if it does not include uptake of the publics' concerns. That is the main reason why transparency cannot be sufficient for the accountability to the publics and for ensuring institutional trustworthiness.

3.4 Conclusion

Based on the results of the analysis of various concepts of trust conducted in the previous chapter, our intention in this second chapter was to articulate a new model of trustworthy institutions that would be applicable in the newly predominant context of strong thin trust in institutions. We formulate the model of trustworthy institutions within typically used three-part relation in which B is trusted by A for specific thing X. In elaborating the model we focus on B as a trustworthy institution that holds properties of competence, responsibility, and responsiveness regarding the specific thing X.

Onora O'Neill's (2013) conceptual framework that suggests the possibility of placing trust well only if we trust the trustworthy is the starting point in formulating the model that prioritizes trustworthiness rather than more commonly used models of trust. We find it a more distinctive approach for articulating properties and related mechanisms of institutional design that could be reflected upon when placing trust wisely. We extend this initial framework, however, by suggesting that the role that values, emotions, and affective attitude can play in the context of trust and trustworthiness should not be dismissed, but should neither be given priority, as that could lead to blindly placing trust.

We distinguish two intrinsic aspects of the concept of trustworthiness intentional and normative aspect of placing trust. Next to criteria of competence that has been adopted across various concepts of trust, we extend the formulation of trustworthiness in institutions introducing responsibility and responsiveness as two distinct properties of the institutional design. They are relevant primarily as a means of incentivizing trustworthiness that should be judged in order to place our trust wisely. By introducing properties of responsibility and responsiveness we intend to satisfy the requirement of trustworthiness both on the part of the trustee as on the part of the trustor. In the context of science and scientific institutions, it would mean that public trust in science and scientific institutions simultaneously requires trust in publics.

Our analysis of the mechanisms of openness and accountability starts with O'Neill's (2014) proposal of their revised version. Instead of the managerial and micromanaging form of accountability, we propose it to be implemented based on the meso and macro level that simultaneously enables institutional learning. Analysis of the mechanisms of transparency and

openness indicates the potential for misuse and deception, while O'Neill's suggestion offers a better prospect of introducing intelligible, accessible and assessable communication. This kind of communication moreover satisfies the criteria of trustworthy science when conducted *with* publics rather than *to* publics and taking publics concerns into account. Instead of only signaling it also includes uptake of publics concerns in the form of responsiveness.

Concludingly, trustworthy institutions should not only signal its trustworthiness to the trustor but also be responsive in a way to include uptake of trustees potential concerns and values regarding the specific thing in question. That is the main extension of our model of trustworthiness that requires from trustworthy institutions not only to signal to trustor their trustworthiness but also to include responsiveness in a form of uptake of concerns and values that A holds regarding specific thing X. New technological advances are already offering new technological means of implementing this model of trustworthiness in the very design of the decentralized systems such as blockchain technologies that have been examined in the first chapter.

CHAPTER 4: MAPPING THE INDICATORS THAT INFLUENCE PUBLIC MISTRUST IN SCIENCE

4.1 Scientific Objectivity

The aim of the final part of the research will be to identify the conditions needed for scientific knowledge and expertise to be trustworthy. Our hypothesis is that in devising the scientific advice both epistemic and non-epistemic, social conditions of trustworthiness have to be met. We argue that in using the scientific advice we have to be aware of its twofold nature. Not only does it refer to high credibility of scientific knowledge, its reliability and sufficient depth, but it also relates to the values, broader political and social context and public contribution to the policy advice. This section will join our research on trust in previous two chapters with discussion on the scientific objectivity and how it relates to the notion of values in science.

Our research on trust so far provided review and analysis of the extent literature and the state of the art in the research on trust, various philosophical dimensions of trust, and the difference between trust and trustworthiness. We focused on the aspects of trustworthiness that received relatively little attention in the scientific literature to demonstrate its significance in the current debate about the science-policy interface. The conceptual framework on the difference between trust and trustworthiness was primarily based on the work of Onora O'Neil's insight on the concept of trustworthiness further extended to elaboration of the model of trustworthiness by design. Its intention is to set a broader framework for possible applications of the ideas to the practical issues of trustworthiness.

These perspectives on trust and trustworthiness will be brought together in correlating scientific expertise with trustworthiness in order to single out the conditions that are needed for scientific knowledge and expertise to be trustworthy. Our hypothesis is that in devising scientific advice both epistemic and social conditions of trustworthiness have to be met. In order to adequately clarify the distinction and elaborate epistemic and non-epistemic values in deciding on trustworthiness of scientific expertise we will examine the research on scientific objectivity and the role of values in science.

First, we will give a short overview of the different concepts of scientific objectivity such as faithfulness to facts, freedom from personal biases and focus on the third concept of value-free

ideal of science and the question of its autonomy. We will tackle the current discussion in philosophy of science about value-free ideal of science that in order to be objective has to be free of moral, political and social values.

Reiss and Sprenger (2017) distinguish between three concepts of scientific objectivity, faithfulness to facts, freedom from personal biases and value-free ideal which we will further address in this chapter. The approach to scientific objectivity in all three concepts is based on the scientific reasoning in terms of theory choice, inference and experiments. Reiss and Sprenger (2017) emphasise two aspects of scientific objectivity. First is product objectivity where products of science, scientific theories, results of experiments and scientific laws accurately represent reality without human influence. Process objectivity is another aspect of scientific objectivity that assumes scientific methods, measurements and processes to be free of social influence or individual biases.

The first concept of scientific objectivity should assure that scientific claims are faithful to the facts and that they faithfully describe world around us. This approach has been elaborated by Popper, Carnap, Hempel, and Reichenbach and is very close to the scientific realism. Scientific method should assure that scientific evidence speaks in favour of scientific claim. However, that is not so obvious since several challenges have to be addressed in relation between scientific evidence and hypothesis in terms of theory-leadenness and incommensurability (Kuhn, 1962 [1970], Feyerabend, 1962).

Faithfulness to facts approach was challenged as unattainable in form of view from nowhere – where processes would not be influenced by human values or goals. Popper (1972, 1934 [2002]) view on objectivity differs from faithfulness of scientific claims to fact. Instead his notion of objectivity relies on inter-subjectively testing and criticism "the objectivity of scientific statements lies in the fact that they can be inter-subjectively tested" (Popper 1934 [2002], p. 22). This line of argument Longino (1990) reinforces in her view of science as a social knowledge rather than faithfulness to facts. She introduces contextual empiricism that assures objectivity of science by transformative criticism where diverse voices and perspectives are heard. Her transformative criticism in interactions between scientists has to be conducted within avenues for criticism enabled by scientific institutions, have to comply to same standards, ensure uptake of criticism and share equality of intellectual authority.

The second approach formulates scientific objectivity as the value-free ideal ensuring that scientific claims are not influenced by political social or moral values. And finally, the third

approach of scientific objectivity is freedom of personal biases that should assure individual preferences and perspectives have no influence on scientific results. Measurement and quantification should be in place instead (Reiss and Sprenger, 2017). In the following chapter we will address those two ideals in more details and within two particular applications in the context of trust and trustworthiness. First in the form of trust in individual experts and second in form of institutional distrust within the broader context of society and policy.

However, before we start that analysis we should not omit another curtail aspect in approach to scientific objectivity. It relates to society and policy and the attributes that give the objectivity of science its rightful place and trust in society. To address this aspect of science and its relation to society and policy we have to take into consideration the social organization of science and how society influences it. Another benefit of scientific objectivity being a basis for the policy decisions and advice.

Within the social constructivism the role of society in science and in particular in application-oriented research becomes more prominent. David Bloor (1976) emphasized social influence on science and on scientists when it comes to assessing hypothesis (Bloor, 1976, p. 873). Carrier (2016) introduces aspects of the social influence on organization of science such as economics benefits and certain interests that are sometimes pushed by politics. They can be evident in terms of setting scientific agenda, requiring secrecy of some research and intellectual property rights that might collide with open scientific discourse.

Another angle of the social constructivism and the role of society in science is evident in inclusion of local knowledge expanding the contributions to scientific discourse to so called “lay experts” or “experience-based experts” (Carrier, 2016). Inclusion of this type of knowledge proved to be of particular importance when science is needed for policy advice on particular topic or domain where local experience-based experts might have invaluable insights due to their familiarity with the issue at hand and local insight.

Brian Wynne (1996) case study gives in particular valuable insights on the value of expertise of laypeople who have local experience-based expertise in particular domain. This case study shows how British sheep farmers who had long term experience with radioactivity from Windscale plant that was near by had lay knowledge that could be aplicable to the issues hey have encountered after Chernobyl accident in 1986. However the lay expertise of sheep farmers was diregarded by scientists and non-epistemic considerations were no taken into account

although in that particular problem at hand it offered invalueble insights. Carrier (2010) ppoints out the incapacity of scientists "to adjust their general, science-based models to the local circumstances and failed completely" (Carrier, 2010, p. 200). In terms of addressing specific challenges Carrier (2010) emphasises the necessity of including different disciplines as well as experience-based knowledge. This aspect of social conditions of trustworthiness is highly relevant in terms of challenges that lay in the the nexus between science and society, "the ability to take up social hopes and fears, or aspirations and concerns, is an essential element of good expert advice." (Carrier, 2010, p. 206).

These social processes and participation of the public should deserve similar significance when we address trustworthiness in the context of science and scientific expertise as other two epistemic and social preconditions that Carrier (2010) proposes for the trustworthiness of scientific expertise, mainly credibility and significance of scientific knowledge. In the final part of our proposal of trustworthy scientific institutions in the Chapter 5 we will elaborate the aspects of social participation within the principle of responsibility, responsiveness and democratic accountability.

4.1.1 Science values and trust

In this chapter we will focus on the third concept of scientific objectivity, namely Value-Free Ideal, its attainability and the roles for values in science. We will apply it by examining the trustworthiness of scientific institutions that has greater promise in addressing the question of trust and mistrust in scientific context than more common appeals to trust in individual experts or within institutional skepticism approach. Scientific expertise and knowledge contribute to decision-making processes and have an impact on the society in many complex issues from biomedical, genomic research to climate change.

However, research controversies make it evident that together with benefits there are possible risks for society implied in epistemic dependence. Due to the lack of capacity to verify each of scientific claims (such as the earth is round) we often have to rely for our knowledge on the testimony of scientific experts and others (Hardwig, 1991). Experts might disagree and scientific claims contested. Confirmation biases, potential unwarranted consequences of

politicization and commercialization of science all might call in question trust in science and be possible causes of mistrust among public.

We will compare two ways of addressing this challenge: (1) trust in individual experts based on the account of individual expert responsibility (Douglas, 2008, 2009, 2015) and (2) institutional distrust approach - based on the institutionally organized skepticism (Merton, 1938; Bouchard, 2016) and transformative criticism (Longino, 1990, 2002). Our line of argument, however, departs from the prospect of promoting trust in individual expertise because it might fail to address the question of how to place trust wisely, while institutional distrust approach does not sufficiently acknowledge the conditions of the current societal context in which science takes place.

In order to better respond to these objections, we propose a third approach that utilizes the concept of trustworthy scientific institutions. This enables us to apply a more selective approach to trust by suggesting criteria within an institutional context that should be met when singling out relevant properties that would make scientific knowledge and scientific expertise trustworthy. We propose an analysis of the principles of integrity, responsibility, democratic accountability and openness in determining their relevance to the question of trustworthiness. Our analysis uncovers various forms of underlying ignorance, either as exclusion or active production of doubt and raises several questions related to the objectivity of science. We argue that trustworthy science requires reformulation of the notion of objectivity in science to ensure responsibility and democratic accountability.

4.1.2 Trust and trustworthiness

Before presenting three approaches regarding trust, distrust, and trustworthiness in the context of science, we will begin by briefly summarizing their meaning in terms of how they will be used based on the analysis in the previous chapter. We understand trust and trustworthiness as a cluster of concepts (Hosking, 2014) that are very often used in a variety of meanings. Research on trust in moral philosophy has predominantly been done with regards to trust between individual agents and mainly viewed as motivated by the good will towards the trustor (Jones, 1996; Baier, 1986). Those accounts, however, acknowledge a risky aspect of trusting relation as

trust can be betrayed and cannot answer the question of how to place trust wisely. Therefore, we change the focus from the trust to trustworthiness, as the trust is well placed in someone only if the trustee is trustworthy.

Distrust can be seen as opposite to trust. The view of trust as a virtue assumes trust to be moral, thus implying that distrust should be bad. But distrust is sometimes good as it might protect us from damaging consequences of misplaced trust. Hardin gives an example of parents who do not entrust the safety of their child to unworthy caretakers, to demonstrate how distrust can be good (Hardin, 2002). In the context of scientific criticism - as we will see in the next section - distrust can also have beneficial implications in uncovering biases and background assumptions and thus can contribute to scientific objectivity.

In our approach, we focus on the concept of trustworthiness as we are interested in determining how to achieve more trust to be well-placed in the trustworthy. Onora O'Neill frames this challenge of intelligently placing trust: "If blanket scepticism is not a feasible basis for life we must place trust selectively and with discrimination even when we lack any guarantee that agents or institutions of any specific sort are unfailingly trustworthy. The possibility of being mistaken, deceived and even betrayed cannot be written out of life. It is therefore important to find at least approximate ways of distinguishing between well-placed trust and misplaced trust" (O'Neill, 2002a, p. 122-123). In order to place our trust wisely, we extend her proposal to trustworthy institutional model that should assist in judging competence, responsibility, and responsiveness of institutions related to a specific purpose at hand.

At this point, our intention is to apply the model of trustworthy institutions to the scientific institutional setting of trustworthy scientific institutions. Onora O'Neill (2002a) in this regards emphasizes that experts have highly specialized knowledge in science, medicine etc. that is not easily judged by citizens. Moreover, institutions - including science - are becoming more complex and obscure. In this constellation, it is much harder for ordinary individual or citizen to judge where to place trust and for what specific purpose. It is much more challenging to judge whether to place trust in complex institutions than in individuals, how to properly assess evidence or institutional safeguards (Weinstock, 2013).

Our analysis in the first chapter exposed potential flaws in the encapsulated interest concept of trust as well as in the proposal of substituting trust with cooperation (Cook et al., 2005). Nevertheless, the distinction between trust and trustworthiness as well as the broader

framework of internal, external and mixed inducements for trust elaborated by Hardin still gives a relevant framework for our concept of trustworthy institutions. Hardin describes trustworthiness as our commitment to fulfil the trust that someone put in us, “trustworthiness is a motivation or set of motivations for acting" (Hardin, 2002, p. 31).

Hardin proposes three categories of reasons to fulfil a commitment of another's trust in us. They are internal inducements or external inducements such as legal and other institutional constraints. The final category is a mixed inducement or a combination of internal and external inducements such as norms that can also be internalized in a way that we act upon them even though we will not be sanctioned. He further maintains “creating institutions that help secure trustworthiness thus help to support or induce trust” (Hardin, 2002, p. 30) and he gives a particular attention to discussing cases of mistakenly speaking of trust while the issue is in fact about trustworthiness.

The main distinction between trust and trustworthiness is that trust is an attitude, while trustworthiness is a property. Trust is well placed only if we trust the trustworthy. Hence, introducing the concept of institutional trustworthiness will enable us to adopt a more selective approach to trust by suggesting criteria that should be met when singling out relevant properties that would make scientific expertise trustworthy.

4.1.3 Science and values

The issue of trustworthiness of scientific institutions is placed within the broader context of the role of values in science – a multifaceted question that provokes dissent views. Almost everyone would agree that values guide scientists when they choose adequate methodology in dealing with laboratory animals or human subjects, or when they decide on a research topic or research application. However, the debate on the role of values heats up when it comes to the very core of the scientific process of justification.

The debate is mainly between proponents and opponents of the value-free ideal of science that considers scientific objectivity free of moral, political and social values. It gathers on one side several proponents of the value-free Ideal in science that is seen as a mean of excluding individual and institutional interests and values (McMullin, 1982; Lacey, 1999; Mitchell, 2004).

The opposition gathers critics of the value-free ideal of science (Longino, 1990; Douglas, 2009; Kitcher, 2011).

The value-free ideal of science is also considered one of the three concepts of scientific objectivity, together with the concept of faithfulness to facts and freedom from personal biases (Reiss and Sprenger, 2017). The proponents of the value-free ideal of science argue it should be free from various moral, social and political values and should eliminate value judgments from the core of the scientific process of justification in order to protect the objectivity of science (Lacey, 1999; McMullin, 1982; Mitchell, 2004). The on-going discussion, therefore, focuses mainly on two central stages in scientific cycle - gathering evidence and the acceptance of scientific theories – and questions if they are value-free and should they be value-free. The first and last stage in the scientific cycles – on the choice of research problem and the application of the scientific research (Reiss and Sprenger, 2017) - are mostly not included in the discussion, as they are already perceived as influenced by the values of individual researchers and societies, although as it will be later elaborated, some does not take this distinction to be so obvious.

Furthermore, it is necessary to outline the distinction between the epistemic (cognitive values) and contextual values (non-cognitive values) as the value-free ideal in science considers that only the former are problematic. Consequently, it might be argued that the term value-free ideal might also be problematic as it does not imply the exclusion of epistemic values and thus does not refer to all values.

But at this point, we are interested mainly in the distinction between different types of values. Epistemic or cognitive values are first called values in the Thomas Kuhn's paper "Objectivity, Value Judgement, and Theory Choice" and are understood as "accuracy, consistency, scope, simplicity and fruitfulness" (Kuhn, 1977, p. 322). They are then further elaborated and diversified in different terminology formulations in the work of Laudan (2004), Douglas (2013), Lycan (1985) and McMullin (2013).

The value-free ideal focuses solely on epistemic values in the scientific process and it is consequently directed to reducing the contextual values such as moral, social, political, personal or cultural values, because according to this ideal, they should not have an impact on scientific reasoning (Reiss and Sprenger 2017; Dorato, 2004; Ruphy, 2006; Biddle, 2013). The value-neutrality thesis, on the other hand, investigates whether it is possible to achieve the value-free

ideal (Reiss and Sprenger, 2017) and it is related to the difference between the context of discovery and on the other side the context of justification (Reiss and Sprenger, 2017).

On contrary, the value-laden thesis argues that both epistemic, as well as contextual values, are crucial in any scientific endeavor; that we should abandon the value-free ideal (Feyerabend, 1978) and that we should redefine the ideal of objectivity (Longino, 1990; Douglas, 2009). Merely excluding all values from the scientific process of justification, they argue, does not show that the research process is, in fact, free of values nor does it prove that values cannot also play a positive role (Longino, 1990, 2002; Douglas, 2000, 2009; Kitcher, 2011). The further discussion relates to the debate if the distinction between epistemic and non-epistemic values is even possible (Longino, 1996). And if the distinction is possible, should epistemic values have priority over non-epistemic values in terms of obtaining the truth about the world, or other goals of scientific research should have equal standing (Rooney, 2017).

Within the value-laden thesis, there are several lines of arguments regarding the possible places for the positive role of non-epistemic values in science. Among them, three might be singled out as the most relevant. First place for values is introduced in the context of the feminist critique of science and based on the underdetermination argument (Harding, 1991; Longino, 1990, 2002). The argument assumes that science is undetermined by evidence which opens the gap for the influence of values. The second challenge to the value-free ideal of science is articulated in inductive risk argument (Rudner, 1953; Churchman, 1948) and further elaborated in the work of Heather Douglas (Douglas, 2000, 2009).

It shows that scientists faced with the inductive risk, in fact, employ value judgments. Namely, they have to decide whether there is enough evidence for the claim depending on the possible consequences of false positive or false negatives. Janet Kourany (Kourany, 2010) articulates third possible positive places for values in science articulating it in the ideal of socially responsible science based on the argument for the right values in science and importance of epistemic as well as social standards.

Furthermore, within this second challenge articulated within the inductive risk argument, Douglas emphasize that values play a legitimate role in expert reasoning only in their indirect role in determining the acceptable level of uncertainty and the evidential threshold for the claim in particular case. Therefore Douglas (Douglas, 2000, 2009) introduces differentiation also between indirect and direct roles of values as a crucial argument in preserving the integrity of science.

According to Douglas, contextual values (under which she counts not only ethical and social but also cognitive values) should play indirect role in the so-called "internal" part of the research process when researchers decide on the evidential sufficiency for the acceptance of the hypothesis bearing in mind the consequences of false positive and false negative as well as in the interpretation of the evidence. Values should not play a direct role in this internal part of the scientific process, because if that happens then hypothesis would be accepted based on our values and preferences instead of being based on evidence. That would allow rejecting or adopting hypothesis according to a predetermined outcome.

She holds values legitimate in their direct role influencing choices only in the so-called external part of the scientific process; in the choice of the research subject, methodology (that might mean refusing to use a specific methodology that would cause harm to animals, e.g.) and the choice of application of the research. Douglas (2009, 2013) distinguishes the indirect role of contextual values from the role of epistemic values whose role is to assist us in determining the level of uncertainty, while contextual values then decide if that level of uncertainty is acceptable or not.

However, her approach to values in their direct or indirect role if cantered on the trust in an individual scientist, might not sufficiently take into account the broader social context in which science is conducted. Although ethical guidelines for individual scientists might be helpful in distinguishing between direct roles of values (that directly influence the choice of the theory) and indirect roles of values (in evaluating sufficient evidence of accepting the theory based on its possible consequences), it may also be questioned if that formula is always practically achievable.

The choice of the methodological approach scientist want to apply to specific research question surely would not be contested. However, the responsibility posed on each individual scientist in applying indirect roles of values in such a constellation might be more problematic. Furthermore, it might also be questionable if this ideal is attainable, are scientists aware of each part of the process where they should apply values in indirect roles and is performing policing on individual scientists in practical terms attainable or even needed.

As a solution Douglas (2006) proposes that we should follow the way scientist answers to the new evidence or to the critic. Whether he is changing his claims after that or not, should then be our clue if a scientist is using values in a direct role or holds strong values. That should enable us to decide should we trust that a specific scientist or not. Also, according to Douglas (2015) if

we do not share the same views on the way specific scientist is weighing the risks of accepting or not accepting hypothesis in the context of possible consequences of false positive or false negative, we should reject the results.

This proposed solution of post-fest deciding whether we agree with scientist or not does not seem entirely convincing. Furthermore, rather than functioning only in his capacity as individual, scientist also operates in the context of a scientific community that sets standards of conduct in different levels of research as a part of its practice, that should also not be ignored.

Finally, it is also necessary to point out that distinguishing different parts of the research process, namely internal and external parts, rather than being beneficial, might undermine the discussion on values and their proper role. Discussion in the philosophy of science between proponents and opponents of the value-free ideal of science that considers scientific objectivity to be free of moral, political and social values focuses mainly on "internal" stages in the scientific cycle when scientists gather, interpret evidence and accept scientific theories. "External" parts of scientific cycles related to what research do scientists pursue and how will the science be used and applied (Reiss and Sprenger 2017) - are mostly not included in the discussion, as they are already perceived as influenced by the values of individual researchers and societies.

Focusing only on the internal part, however, fails to recognize that the acceptance of the hypothesis in the internal part of the research process is also influenced by the external part (Okruhlik, 1994; Kourany, 2010; Brigandt, 2015) because only theories that are available in the first place could be tested and adopted. This distinction, therefore, might overlook the implications that decisions on research agenda might exert on the research itself, or so-called "internal" parts of the scientific process. Focusing only on the "internal" part of the scientific process might result in ignorance about possible macro-biases in setting the research direction and agenda as will be and further examined within the concept of institutional distrust.

4.2 Trust in individual experts or institutional distrust

4.2.1 Trust in individual expert

The first approach explores possible ways of rebuilding and reinforcing public trust in individual experts. It is based on the moral requirements of the experts due to their responsibility for potential consequences that scientific claims might have on society. Due to societal consequences of scientific expertise Douglas (2008) emphasizes the responsibilities of scientific experts both in seeking the truth and responsible empirical judgments when dealing with uncertainties. The responsibility of experts makes ethical values necessary as moral obligations similar to ones that apply to all of us regarding the consequences of our actions preventing us to be either reckless or negligent.

However, Douglas emphasizes that values play a legitimate role in expert reasoning only in their indirect role in determining an acceptable level of uncertainty and the evidential threshold for the claim in a particular case. In their direct role as wishful thinking values are unacceptable. Therefore, scientific experts are morally required to use social or ethical value judgments in their indirect role but in open and explicit ways in order to comply with democratic accountability.

Douglas indicates several aspects that might influence trust in experts. Openly presenting value judgments should enable publics and policymakers to determine whether to trust experts, they should not be trusted if they use values in direct role, do not respond to criticism and do not change their view when presented with new evidence. That should be an indicator that they do not reason with integrity. Another aspect that might influence trust in experts (Douglas, 2015) is to determine whether they weigh the evidence in the same way we would and, thus, share the same social and ethical values with us.

Although this account of tracing values behind experts' claims should serve as guidelines to help the publics in deciding if they agree with experts and in the final instance ensure trust in individual experts, we argue that doing so might not fulfil its purpose. This complex procedure might make it very difficult for non-experts to navigate. Even if they manage to do so, is it realistic to expect experts to be open about all values that are guiding their judgments and furthermore could values behind their reasoning always be evident to experts themselves? Those

aspects of unintelligibility and *post festum* tracing values could hinder the prospect of publics' trust in scientific expertise.

De Melo-Martin and Intemann (de Melo-Martín and Intemann, 2016, p. 512) criticized this approach since it would be hardly attainable to "backtrack" value decisions and assess possible outcomes if different value judgments were employed. It would require specific expertise, and even then the unpredictability of conclusions of other value judgments would pose another challenge. Even if it might be possible, they argue, what would be the course of action if there were a disagreement about the value judgments made along the process. Should policymakers disregard the research results and made uninformed policy decisions or precede according to research results although they are not in line with democratic values?

There are, nevertheless, several advantages of Douglas's proposal regarding the principle of openness and transparency. The requirement to be transparent and explicit about value judgments contributes to democratic accountability of trustworthy science as will be elaborated in the final section. Furthermore, openly presenting value judgments (Douglas, 2015) subjects experts to critique from various perspectives that is the basis of both transformative criticism (Longino, 1990, 2002) and organized skepticism (Merton, 1938) that will be examined in the second approach of distrustful institutions.

4.2.2 Institutional distrust

The second approach emphasizes that trust should not be placed in individual experts but rather in science as an institution that is a system based on the critical and skeptical approach to expertise, as suggested by Bouchard (2016). It does not assume that trust always has to be good, especially if it might have dangerous consequences. Distrust in individual experts, on the other hand, in terms of critical approach, paradoxically might be one of the aspects that contribute to trust in science as an institution. Criticism should keep possible individual biases in check ensuring objectivity on the level of scientific community and enable scientific expertise to provide our most reliable knowledge. This notion of institutional distrust will be based on the premises of the uncertain nature of science (Douglas, 2015), norms of Merton's organized skepticism and Longino's transformative criticism.

Merton's introduction of "ethos of science" indicates that next to epistemic authority of scientific expertise, certain values might produce better and more reliable knowledge when guiding knowledge production within scientific institutions. Among the four main norms of science that Merton (Merton, 1938) introduces are universalism, communalism, disinterestedness, and organized skepticism. Universalism means that the acceptance of claims does not depend on the personal or social characteristic of scientist, such as race, nationality, class, religion. Communalism, earlier called communism, attributes scientific knowledge as a common property, not kept in secrecy, but openly communicated. As such, it thus contributes to the cumulative character of knowledge as encapsulated in Newton's sentence "If I have seen farther it is by standing on the shoulders of giants". The last two are the most relevant for the formulation of the concept of institutional distrust. Namely, the disinterestedness that is achieved in the institutional scientific setting where experts are under scrutiny by each other ensuring the integrity of science and organized skepticism that enables scrutinizing individual beliefs in the scientific institutional setting.

This final norm of organized skepticism can be seen as further elaborated in Longino's articulation of the conditions of transformative criticism (1) public venues for criticism (such as journals and conferences), (2) uptake of criticism that might influence modification of beliefs or further develop arguments, (3) public standards, shared values and standards for evaluation and (4) tempered equality of intellectual authority that should be granted to all participants (Longino, 2002, p. 129-133). Longino (1990, 2002) thus proposes to widen scientific norms by four norms for social structures of epistemic community. To transform subjectivity into objectivity, she argues, the process of justification must be social in order to tackle background assumptions and biases of individuals and subgroups.

Furthermore, this notion is also at the very heart of uncertain nature of science that according to Douglas by its openness to critical scrutiny ensures that science is our most reliable source of knowledge (Douglas, 2015, p. 301). This nuanced understanding of uncertain nature of science might justify placing trust in science as an institution because of distrustful relation and disagreements between scientific experts. However, we will argue that this critical approach might also entail several disadvantages as it might be misused for achieving specific economic or political interest as will be illustrated based on the Tobacco industry case study.

4.3 Tobacco industry case study

The model of institutional distrust, although rightly grounded on the premises of organized skepticism within scientific institutions, nevertheless fails to acknowledge the broader societal context in which science takes place and as such might not adequately address the whole spectrum of trusting relations in regards to scientific institutions. In order to clarify this claim, we will use tobacco industry case study as a basis for distinguishing four aspects of the embeddedness of scientific institutions in the broader societal context. This interconnectedness consequently impacts public health and is of importance for the decision-making processes.

Critical points of the tobacco industry case study might be best distinguished based on the research that Robert Proctor (Proctor, 1995, 1996, 1999, 2008, 2011) conducted on the tobacco industry and its impact on health. He uncovered different mechanisms that were put in place in order to control science and create ignorance. His analysis starts in the 1950s when the hazardous effects of cigarettes on human health were revealed. At that point, the reaction on behalf of the tobacco industry was strategically orchestrated and enhanced from merely marketing to developing mechanisms to control science.

According to Proctor, the strategy was enfolding in three main ways. First, by ignoring the hazards and referring to the epidemiology as mere statistics. Secondly, by calling for more proof because the evidence of experiments on animals was not good enough for drawing the conclusions for humans. The standard for evidence was put so high that no evidence or proof was enough and further evidence was needed. Thirdly, it was by funding research that should provide predetermined results, namely that the causes of cancer are not known. All those were the ways that manifested the intentions to control science. Ignorance and doubt were manufactured via the establishment of Tobacco industry research council, by lobbying and predetermining research results by the Tobacco Institute and by spreading misinformation via funded projects, publications, press releases or adds in popular scientific magazines.

Based on the tobacco industry example and in relation to broader societal context, we will focus on four critical points. First two indicate possible misuse of uncertainties and openness for delaying the regulations: (1) misuse of the principle of openness and transparency (2) high standards of evidence for the policy decisions that might be set in order to delay regulations. Another two points indicate biases that could influence changes in trust by the publics: (3) active

and passive ignorance that could be reflected in the form of macro-biases and funding of science and finally, (4) confirmation biases that contribute to wishful thinking. All those aspects, we will argue, have direct implications for the embeddedness of science in societal context as well as on influencing changes in public trust. Therefore, in the final instance, they play a significant role for the prospects of trustworthy scientific institutions. This particular tobacco industry case study was chosen since it uncovers mechanisms that were first employed within the tobacco strategy, later to be found and implemented in various different commercial, industry and societal contexts.

4.3.1 Misusing uncertainties and openness

4.3.1.1 Mechanism of openness and its misuse

For the purpose of illustrating (1) first point related to the misuse of the principles of openness and transparency, we will point out to one specific aspect of the tobacco industry case study on critical approach within uncertain nature of science that has recently been addressed by Levy and Johns (Levy and Johns, 2016). They examine two aspects of "institutionalization of uncertainty" through laws that are regulating transparency of publicly funded science. They show how industry-backed "sound science " team in the Philip Morris company in the 1990s influenced the introduction of two acts that are regulating openness of scientific information and enabling their criticism.

One of the introduced acts is the Data Access Act (DAA), which requires public access to publicly funded science but is not applied to privately funded science. The second act Data Quality Act (DQA) should enable "affected persons" to seek correction of the information. After the acts were introduced, the closer analysis showed that criticism of openly available scientific results mainly tended to be used by the industry, lobbyist and trade organizations (Levy and Johns, 2016). Misusing an appeal to the uncertain nature of science enabled obstructing inconvenient regulations and advancing political interests.

This example indicates how openness in science and governance, next to benefits in terms of accountability can also serve as a "Trojan Horse" through which other interests can be

pursued. Openness and transparency of the uncertain nature of scientific results might misuse organized skepticism for specific political and commercial interests. Manufacturing doubt in this "Trojan horse" example uses the same norm of criticism enabled by the open science and open data, not motivated by reaching the truth or ferreting illegitimate values, but for delaying regulations. It thus might have consequences for the public and contribute to public mistrust. On contrary, the same critical approach cannot be applied to some of the industrial research results due to its secrecy that therefore hinders critical scrutiny by other experts and organized skepticism in a Mertonian sense.

As we have previously indicated, the tobacco industry case is not only related to misuse of openness and transparency. It offers an insight into several aspects of ignorance in its active and passive role, but also to an array of biases ranging from confirmation to macro biases. As such, this case study offers valuable insights to ignorance that underlies the intersection between science and values. Therefore, before taking a task of formulating a proposal of the model of trustworthy scientific institutions, we should first further examine next three points that place scientific research in broader societal context within tobacco industry case study.

(1) High standard of evidence.

Uncertainties are an integral aspect of the scientific enterprise. Douglas (2017) emphasizes uncertainty in science due to its inductive method. There can never be complete evidence for accepting the hypothesis so scientists have to use value judgements to determine the evidential threshold. The research on high standards of evidentiary threshold and their implications has initiated a substantial research interest (Proctor, 2008, 2012; Wilholt, 2009, 2013; Elliott and Resnik, 2014; John, 2015; Steel, 2016). It indicates that misuse of high standards of evidence potentially can have a significant impact on the perceived objectivity of science. Furthermore, although surveys on public trust in science suggest that trust is generally high, when looking closer into its potential risks and dangers the view becomes somewhat different. Financial dependence of science and technology is of a concern for 58% of Europeans according to the Eurobarometer Survey dedicated to science and technology (Eurobarometer Special Survey, 2010, p. 19).

By introducing two case studies we intend to further examine how can setting high evidentiary standards potentially impact objectivity of science and what is the role of financial

dependence financial biases and conflict of interests in influencing public trust in science. Examining two case studies - on the tobacco industry and endocrine disrupting chemicals - within two interconnected perspectives will provide further insights into possible indicators of changes in public trust in science.

In the tobacco industry case study, Proctor (2008, 2011) uncovers ways in which uncertain nature of science can be misused for achieving specific economic or political interests. Hidden interests behind the tobacco industry misused scientific uncertainty by calling for ever more evidence to prove hazardous effects of cigarettes. The standard for evidence was put so high that no evidence or proof was enough and further evidence was always needed. Scientific results were challenged by ignoring the hazards, disregarding epidemiology studies and calling for ever more proof before the conclusions could be drown.

> "'More research' is always needed, a 'benefit of the doubt' is always granted, as if cigarettes were on trial and innocent until proven guilty. The industry loves this form of the 'null hypothesis': they start by assuming 'no harm done', and then fail in their feeble efforts at falsification. Similar strategies have been used by other industries to disapprove hazards of lead, asbestos, and the like: and petrochemical and neoconservative doubters of global warming have learned a lesson or two from the tobacco doubt mongers" (Proctor and Schiebinger, 2008, p. 18).

Setting high evidential sufficiency standard before accepting hypothesis - as illustrated in the tobacco industry case study - can be exploited in order to delay unwanted regulations and policy decisions. As long as the standard of proof is set so high, the doubt will be maintained and there will be more need for research.

The second case study on endocrine disrupting chemicals will enable us to examine further implications that setting high standards of evidence might have in the interplay between science, policy, society and public trust. The debate on how should the standard of evidence be determined for the policy decisions was triggered by the European Commission (EC) regulation report on the endocrine disruptive chemicals (ECD) (EC 2013; Horel and Bienkowski, 2013). After the report was published the group of journal editors (Dietrich et al., 2013) criticized two main presumptions of the EC report in terms of what are the effects of the endocrine disruptive chemicals and on what evidence can be accepted as sufficient for policy decisions.

First "the authors expressed the concern that the EC would presume that EDCs do not have a threshold dose below which they cease to induce adverse effects" (Elliott and Resnik, 2014, p. 647). And the second point of disagreement was whether experiments on animals can be applied to humans in lack of other evidence. EC report, in line with numerous other results (Bergman et al., 2013, Grandjean and Ozonoff, 2013) suggested accepting the evidence contrary criticism of the acceptance that kind of evidence on the other side (Dietrich et al., 2013).

The dispute over the standard of evidence can be analyzed through the lens of value-laden science and potential hazardous consequences for publics when setting high standards of evidence. Analysis of this issue by Elliott and Resnik (2014) suggests that the

> "dispute over the EC report illustrates how scientists are forced to make value judgments about appropriate standards of evidence when informing public policy. Empirical studies provide further evidence that scientists are unavoidably influenced by a variety of potentially subconscious financial, social, political, and personal interests and values" (Elliott and Resnik, 2014, p. 647).

Apart of pointing to the value-laden nature of the scientific process that will have policy implications, they also warn of consequences that not being open about implicit value judgements can have for the objectivity of science and consequently public trust, concluding:

> "When scientific evidence is inconclusive and major regulatory decisions are at stake, it is unrealistic to think that values can be excluded from scientific reasoning. Thus, efforts to suppress or hide interests or values may actually damage scientific objectivity and public trust, whereas a willingness to bring implicit interests and values into the open may be the best path to promoting good science and policy" (Elliott and Resnik, 2014, p. 647).

Elliott and Resnik take this case study to be an illustration of the value-laden character of science since value judgement was needed to determine the standard of evidence that will be used for policy decisions and consequently potentially pose risk for society. In this case, they conclude, the implications of applying high standards of evidence might pose a great risk for publics exposing them to hazardous influences until further evidence would be available.

This interconnectedness between values and empirical evidence has further implications for public trust in science, since

> "society is likely to be better served when scientists strive to be as transparent as possible about the ways that interests and values may influence their judgment and reasoning,

> while still striving for objectivity. Transparency can promote public trust by helping laypeople understand how both empirical evidence and value assumptions enter into scientific decision making and policy formation." (Elliott and Resnik, 2014, p. 648)

Further implications on public trust are potential financial biases later traced to the critics of the EC regulatory report. It was found that the critics of the report had ties to the industry but they failed to declare the conflict of interest. "An investigative report found that 17 of the 18 authors of the initial editorial by Dietrich et al. (2013) had ties to regulated industries (Horel and Bienkowski, 2013). However, some of the authors disregarded the issue of conflict of interest as irrelevant and as distracting the focus from the real problem, denying also that their relation to the industry could have had any influence on their views (Elliott and Resnik, 2014, p. 648-649). On contrary, evidence from social sciences on group polarization and the evidence of the effects of financial interests on our subconscious, suggest otherwise (Elliott and Resnik, 2014, p. 649).

As both case studies show, high standards of evidence in scientific research, financial interests and conflicts of interest related to public policy regulation can have a serious impact on publics and as such influence changes in public trust in science. Setting high standards that demand ever more evidence is undoubtedly relevant in various scientific areas, including the ones with a direct impact on the public. Applying high standards when new drugs should be released, in most cases would be beneficial for society Although also in this case the standard may vary depending on the currently available drugs for that specific disease and how acute is a need for the drug.

In the case of regulating endocrine disrupting chemicals, however, setting high standards of evidence and demanding ever more evidence before regulation can be introduced, can have harmful consequences for the public. The same conclusion could be drawn from the tobacco industry case study. Acknowledging the fact that cigarettes cause cancer and imposing regulations fifty years after the scientific community agreed on that conclusion caused great harms for the public health. Both case studies showed that misusing uncertain nature of science for requesting high standards of evidence in order to delay regulations while at the same time diminishing the role of financial biases can have a significant impact on society and thus influence changes of public trust in science.

4.3.1.2 Macro-biases and confirmation biases

Further analysis of the tobacco industry practices uncovers a significant role that biases, such as macro-biases and confirmation biases, play in influencing science and potentially public trust in science. Macro-biases were introduced by Proctor within his study on agnotology, or a study of ignorance. Introducing the study on agnotology, Proctor expands analysis of ignorance from its common understanding as a lack of knowledge to other forms of active and passive ignorance. Ignorance can produce doubt in its active as well as in passive role by directing research funding and research priorities.

Both active and passive forms are related to the broader societal context and influence public trust in science. Proctor categorizes different types of ignorance, "ignorance as a *native state* (or resource), ignorance as *lost realm* (or selective choice), and ignorance as a deliberately engineered and *strategic ploy* (or active construct)" (Proctor and Schiebinger, 2008, p. 3). The type of native ignorance is nothing new, it is well known in philosophical as well as in scientific discourse at large. It either characterizes absence of knowledge or refers to virtuous ignorance due to ethical considerations, such as when we choose not to conduct research because of its possible harmful consequences.

The other two concepts of active and passive ignorance, on the other hand, have so far not been comprehensively addressed in the philosophical literature. Yet, it seems that exactly those two aspects of ignorance have great significance in terms of the public trust as they are related to the broader societal context of science. Examining them exposes further inadequacies of the institutional distrust model, preparing the ground for the proposal of a more adequate model.

As previously mentioned, both native and virtuous aspects of ignorance are well known and often used notions of ignorance. The first state of ignorance Proctor calls native because it resembles the infant stage when knowledge is still not acquired. It can likewise be seen as a resource fueling science motor with questions since if there were no further questions there would neither be a need for research.

Ignorance can also have its virtuous side, present in regulations or ethical decisions not to conduct the research. Proctor relates this ethical method to Rawls "veil of ignorance". There are numerous examples of dangerous research where ignorance is preferred over knowledge, such as research on humans or animals that should not be conducted if it might present a risk. The list

further extends to research that should not to be disclosed, such as possible dangerous research on bioweapons, or due to reasons of respecting our right to privacy, or as in the case of refusing to publish archaeological findings that were obtained illegally (Proctor and Schiebinger, 2008).

Other forms of ignorance, however, both in their active and passive role were often not addressed in the philosophical literature. Active production of ignorance Proctor illustrates by two products of the tobacco industry, one product are cigarettes and another is doubt. Tobacco industry actively produced ignorance, deception and doubt

> "...Ignorance is the role of the industry itself in creating ignorance: via advertising, duplicitous press releases, funding of decoy research, establishment of scientific front organisations, manipulation of legislative agendas, organisation of friendly research" (Proctor and Schiebinger, 2008, p. 17).

This agnogenesis of the "black art" of the tobacco industry that produces both cigarettes and doubt (in the form of ignorance, misinformation, and misuse of uncertainty) proved as highly beneficial in fighting knowledge with knowledge.

Passive ignorance, on the other hand, can be illustrated within the scientific research process at the first stage in setting research priorities are defining the research subject. Focusing on one research line while ignoring the other can have an important impact as "research lost is not just research delayed; it can also be forever marked or never recovered" (Proctor and Schiebinger, 2008, p. 7). This question of who should be engaged in setting a research agenda and deciding on which research will be performed tends to have a significant impact on public trust and social dimension of knowledge.

In this particular tobacco industry example, Proctor uncovers the ways in which the direction of research and research agenda were defined. The funding tended to be generously provided via Tobacco Industry Research Council (later renamed Council for Tobacco Research) for the research on all other cancer-related topics except on its relation to tobacco. The investigation on a potential correlation between tobacco as a cause of cancer was eliminated in the start. This kind of biases that can be observed in intentionally directing attention to specific research topic - in this case with the purpose of not revealing the correlation between tobacco and cancer - Proctor introduces as macro biases.

The purpose of macro biases "is to jam the scientific airwaves with true but trivial work, distracting from what is going on more fundamentally. Bias in such cases lies typically not in the

falsification or misrepresentation of research (though both of these occur, as we have seen) but rather in the diversion of attention from one problem to another" (Proctor, 2011, p. 131). Those macro biases are not related to misrepresentation of research or badly done science that could be disregarded as illegitimate according to internal epistemic standards. They are rather related to setting priorities in the domain outside of the strictly internal conduct of scientific research.

This consequently leads to acknowledging the relevance of broader societal context when addressing the question of trustworthy scientific institutions. Pinto's analysis of Proctor's rejection of value-free ideal concludes along the same lines:

> "Instead of focusing exclusively on internal epistemic standards, we need to zoom out. In fact, research that complies with internal epistemic standards might contribute to creating epistemic obstacles at a broader scale, a tactic well known to the tobacco industry, which sponsored good-quality research at top universities on lines of research that diverted attention from the causal link between smoking and cancer. Here, Proctor also seems to distance himself from … defense of independent science: adopting the political perspective entails that science is inevitably embedded in a broader socio-political and economic context, from which no true independence is possible" (Pinto, 2015, p. 305).

Acknowledging that broader societal context and recognizing that biases are present at micro and macro level, opens up the discourse of science and values in internal as well as external scientific contexts that are interconnected as the tobacco case study has proved.

There is one more caveat in tobacco industry research funding, where results were predetermined and known even before the research was conducted. That kind of biases of wishful thinking we could further elaborate as confirmation biases. Proctor and Schiebinger (2008) give an example of the Philip Morris's project called Cosmic that in 1987 received extensive funding for a joint research on the drug history. Scientists and historians were to show the benefits of nicotine within the conducted research.

This example shows the possible ways of controlling science and history within a very similar tactic of confirmation biases. When the results were published, however, the industry behind the funding of the project was not acknowledged. The process was later presented by the industry as a conflict between popular and scientific knowledge, while in fact, the strategy was to infiltrate into the scientific discourse and with more than generous funding to fight knowledge

with knowledge. This mechanism of conducting the research to justify predetermined beliefs and wishful thinking consequently also produces confusion and doubt.

4.4 Indicators that are influencing changes in trust

The analysis of the tobacco industry case study provides a framework to identify several indicators that are influencing changes in trust. Features of misusing uncertainties and openness for delaying regulations as well as biases in form of macro and confirmation biases will provide a starting point in our attempt of formulating indicators that cause a decrease in public trust. Based on the tobacco industry case study as well as two proposed models of trust in individual expert and institutional distrust model, we will single out two main distinctive causes. First looks at (1) framing the problem and research agenda that can sometimes neglect different voices, and; second (2) using science for delaying regulations with potentially harmful consequences for publics and public trust in science. We will elaborate on both proposed indicators as follows.

4.4.1 Framing of the research agenda and neglecting different voices

Separating internal from the external part of science results in the scientific practices that indicate possible changes in public trust in science. Our analysis of the first approach of trust in individual experts suggests that it might be problematic to insist on the clear distinction between "internal" part of scientific process when scientists gather evidence, interpret them and accept scientific theories, from "external" part of the research process when scientists choose the research problem and apply scientific results. Focusing on so-called "internal" part of the scientific process as a separate from the "external" stages creates a division that obscures their essential interconnectedness. Focusing only on the internal part of scientific process ignores the influence that external part of the research process has on the external part (Okruhlik, 1994; Kourany, 2010; Brigandt, 2015) as only proposed hypothesis could be tested in the first place. Further analysis of their interconnectedness provides new insights into potential macro-biases in setting the research agenda and research direction (Proctor and Schiebinger, 2008). That is

relevant also in terms of recognizing the role that publics have in framing the problem and research question (Goldenberg, 2016). All these aspects are indicating possible changes of publics' trust in science.

Secondly, our analysis suggested that macro biases that are present in the form of passive ignorance when research is directed in one way while ignoring the other (Proctor and Schiebinger, 2008) can serious long-term consequences for publics' trust in science. In the tobacco case study, the focus was deliberately diverted from research on a potential link between tobacco and cancer to other possible causes of cancer. Applying research integrity only on the internal part of the scientific process, regardless of its rigor, seems to omit the reality of its embeddedness in broader societal context. Research of highest scientific quality and conducted by excellent scientific rigor and integrity should be expected to ensure public trust in science. However, that is not so obvious because research that was intentionally directed away from possible causes of cancer-related to cigarettes by the Tobacco Institute had harmful effects on public health at the same time excluding relevant voices in directing the research agenda. Both of those aspects could have a significant influence on changes in public trust.

4.4.2 Using science for delaying regulations with harmful consequences for the publics and public trust in science

As shown in the analysis of the "Trojan Horse" case study, misuse of the principles of openness and transparency can influence changes in trust. It happens when organized scepticism is misused for specific political and commercial interests that are profiting from the uncertain nature of scientific results. It can contribute to manufacturing doubt by using norms of criticism not with the aim of ferreting illegitimate values but for delaying regulations, potentially having dangerous consequences for public health as well as causing changes in public trust in science.

Setting high standards of evidence could be used for delaying regulations. Therefore it should not be perceived as value-free (Elliot and Resnik, 2014) as it might hide financial biases and conflicts of interests with implications on scientific objectivity. Demanding more evidence in specific scientific disciplines is rightly used as necessary for high epistemic quality. However, when demanding more evidence and setting high epistemic standards might have a direct impact

on delaying regulations and posing risk for the public, harmful consequences for the public should be prevented.

The damaging integrity of science and influencing public mistrust can surely be prevalent when evidence is not distinguished from wishful thinking in form of confirmation biases. As it was presented in the tobacco industry case study where researchers and historians were advice to research topic with predetermined results of showing benefits of tobacco. Also in this case, not declaring the potential conflict of interest and the source of research funding is further problematic. Harmful consequences for the public, confirmation biases, financial biases and conflict of interest all have a strong impact on changes in public trust in science.

4.5 Conclusion

In this chapter, we examined the concept of trustworthy institutions in the context of scientific institutions. When tackling indicators of public mistrust in science we proposed using the model of trustworthy institutions instead of an approach based on the trust in individual experts or within institutional skepticism approach.

Our line of argument departs from the presumption of promoting trust in individual expertise based on the account of individual expert responsibility (Douglas, 2008, 2009, 2015) because it might fail to address the question of how to place trust wisely. The prospect of tracing values behind experts' claims that should enable publics in deciding if they agree with experts and ensure trust in individual experts doesn't seem to be an optimal solution. It both offers too complex of a procedure for non-experts to navigate and poses unrealistic expectation on experts to be open about all values that are guiding their judgments although during the whole scientific process that might not be possible on each of the steps in the process.

Although the proposal of trust in individual scientists does not seem to offer the best solution when it comes to public trust in science, it does indicate a valuable aspect of openness. Proposal regarding the principle of openness and transparency (Douglas, 2009; Elliott and Resnik, 2014) however, in the form of requirement to be explicit and open about value judgments, however, presents one of the valuable aspects of democratic accountability within the prospect of trustworthy science.

The second approach of addressing public trust in science within the institutional distrust approach can be mainly anticipated from the institutionally organized skepticism (Merton, 1938; Bouchard, 2016) and transformative criticism (Longino, 1990, 2002) conceptualizations. In tackling possible changes of public trust in science this approach, however, does not sufficiently acknowledge the relevant consequences of societal context in which science takes place. Tobacco industry case study illustrates how this critical approach might also misuse uncertainties and openness for delaying regulations or on the other hand employ biases in form of macro and confirmation biases. This analysis, however, also proved beneficial as a starting point for formulating indicators of causes for decrease in public trust.

Analysis of both proposed models of trust in individual expert and institutional distrust model extended to the case study of tobacco industry strategy indicated three main causes for

changes of public trust in science. Those indicators then provided fruitful ground for formulating the model of trustworthy institutions that would suggest overcoming some of the indicated challenges. First indicator related to the very framing of the research problem and research direction as well as the practice of exclusion of various voices. Second, misusing openness, high evidentiary standards and organized skepticism for manufacturing doubt and delaying regulations with potentially dangerous consequences for public health and publics. Based on this mapping exercise of indicators that causes changes in public trust in science in the following chapter we will propose a model that should be able to better address indicated challenges.

CHAPTER 5: TRUSTWORTHY SCIENTIFIC INSTITUTIONS

5.1 Principle of responsibility and responsiveness

In order to address the indicators of changing public trust in science within the broader societal context of science, we introduce the third concept of trustworthy scientific institutions. In this attempt, we anticipate not only academia but also other institutional contexts in which scientists work such as in industry, advising bodies, government (Douglas, 2016). The aim is to adopt a more selective approach to trust by suggesting criteria that should be met when singling out relevant properties that would make scientific institutions trustworthy.

We hypothesize that in this quest both epistemic and social conditions have to be met, focusing not only on so-called "internal" aspects of science but also on "external" part and organizational structure of scientific institutions that tend to be equally relevant for the prospect of trustworthy institutions. In this context, we argue that the model of trustworthy scientific institutions requires reformulation of the notion of objectivity in science to ensure principles of responsibility and responsiveness within democratic accountability and the revised notion of openness.

We argue that the model of distrustful institutions should be extended in order to articulate a stronger case for responsibility and responsiveness as an integral part of the shared public standards that are also accommodating extra-scientific criticism. In order to do so, first, we examine Longino's transformative criticism that acknowledges possible benefits of challenging shared assumptions that might come from outside of the scientific community (Longino, 2002, p. 135). However, we suggest that further articulation of extra-scientific critical contributions would be required. Openness to criticism could be applied not only to scientific community and experts but might extend to the broader public and among others could include scientific experts in commercial research as it was shown in the previous example of tobacco industry case study.

Second, we will focus on the dynamic between two social norms that Longino (1990, 2002) introduces: the uptake of criticism and shared public standard. We further examine the ways of reconciling critical scrutiny that contributes to scientific objectivity with its possible misuse in producing ignorance and widening public distrust. Intemann (2017) raises concern about diversity of values and interests in Longino's transformative criticism pointing out to the unintended consequences that give equal weight to feminist values as to sexist, racial and

commercial values and interests. Although proponents of the approach would argue that not all of those values would necessarily survive critical scrutiny, the model still does not adequately answer to the concerns raised by the social, cultural and economic context in which science is conducted and does not seem to fulfil its purpose of minimizing biases (Pinto, 2014; Elliott and Steel, 2017). Tobacco industry case study is another example where the critical approach was misused to institutionalize and codify uncertainty.

Therefore, we argue that within trustworthy scientific institutions - in this particular context of scientific communities governed by transformative criticism - formulation of shared public standards should be extended to include not only epistemic values and standards within different disciplines but also some social values (Kourany, 2010). More precisely, we propose its extension within the framework of trustworthy institutions model where it might contribute to stronger articulation for responsibility and responsiveness. Integrating responsibility and responsiveness as an integral part of institutional design resembles the importance of both signaling trustworthiness as well as responsiveness to public concerns and values within participatory models. Before articulating our proposal of extension of shared publics' standards, we first provide a broader context of Longino's transformative criticism emphasizing the significance of extra-scientific contributions.

5.1.1 Transformative criticism

This brings us to introducing the context for the trustworthy institutions in the scientific and extra-scientific community. Longino (1990, 2002), in these regards, proposes to widen scientific norms by four norms for social structures that would apply to the epistemic community, due to the notion that the objectivity in science is unlikely attainable on the individual level of values and background assumptions of individual scientists. She suggests that within the social structure of scientific communities and practices, scientific objectivity becomes more attainable through critical interaction within the scientific community.

Longino's position originates from the logical problem of underdetermination that was first introduced by Duhem. Underdetermination primarily poses a problem for the justification because it points out to the semantic gap between theories, models, hypotheses and related data

and evidence. The hypothesis can never be fully determined by the data, data can at the same time be consistent with different and even conflicting hypothesis while hypothesis cannot fully specify the data. All those cases open the gap between hypothesis and data that is filled by the background assumptions. Longino refers to background assumptions as the empirical or metaphysical claims that fill the gap between hypothesis and data. Background assumptions are influenced by various geographical, historical or social contexts and there is no prospect of view from nowhere that could resolve decision on the right values. Therefore, she proposes that those subjective preferences could be uncovered and kept in balance by critical social interaction in the scientific community.

The argument for her normative theory of social knowledge is justified by the very way that the knowledge is produced in social and cognitive practices. In this regard, Longino (2002, p. 129-133) proposes transformative criticism that enables transforming subjectivity into objectivity, and she argues that process of justification must be social in order to tackle background assumptions and biases of individuals and subgroups. Among the conditions of transformative criticism she outlines (1) public venues for criticism of research, such as journals and conferences, (2) uptake of criticism that might influence modification of beliefs or further develop arguments, (3) public standards, shared values and standards for evaluation and (4) tempered equality of intellectual authority.

Here we want to emphasize the fourth norm that Longino introduces. It is tempered equality of intellectual authority that features diversity, the characteristic that will be part of the establishing co-creating practices within the model of trustworthiness. Formulating conceptual basis for the necessity of a diversity of perspectives might be most beneficial in terms of uncovering gender and racial biases. Tempered equality of intellectual authority, not only acknowledges but also accepts critic from a diversity of perspectives and multiple points of view. As epistemic criterion, it is not determined by social, economic or political power and it opposes the exclusion of women and racial minorities. Longino pushes the point further arguing that dissenting views should not only be acknowledged but also cultivated. However, the equality is tempered because it is not given to the members of the community who are not taking up criticism themselves, so they are subjected to disqualifying criterion.

Longino's research examines several studies uncovering racial or sex assumptions in research that were not subjected to criticism. Her feminist account of uncovering background

assumptions, posing new questions and providing alternative explanations and approaches are extensively elaborated in the case study on anthropology related to human development (Longino, 1990). Although evidently social norms show robustness for uncovering background assumptions, there might be several possible aspect that the concept fails to address such as shared standards of evaluation (Intemann and de Melo-Martín, 2014) or the paradox between encouraged diversity of values and exclusion of some values (Holter, 2014), but that discussion extends the scope of this chapter.

Although Longino's transformative criticism anticipated the prospect of possible benefits of challenging shared assumptions from outside scientific community (Longino, 2002, p. 135) we suggest that further extension of that argument is needed within the concept of extra-scientific community contributions. Secondly, the formulation of shared public standards also does not seem to be applicable to the broader societal context of macro or confirmation biases. In the next section, we examine and propose a possible extension of those two points.

5.1.2 Extra-scientific critical contributions

As we have implied critical contributions should not only be expected from the scientific community, but also from experts in a commercial context and the broader public. Therefore, normative principles for social structures that Longino proposes should be further extended to extra-scientific critical contributions. Although Longino also acknowledges possible risks of reinforcing assumptions by the scientific community and therefore proposes them to be challenged not only within the scientific community but also from the extra-scientific community, she does not provide further articulation of the argument.

The prospect of expanding the discussion to public participation and extending Longino's argument to extra-scientific community will be examined within the case study on the relations between archaeology and Indigenous descendant community in North America (Wylie, 2014). Wylie focuses on the 'criticism from multiple points of view' introduced by Longino (2002), arguing that the fourth norm of tempered equality of epistemic authority should be redefined and extend to criticism external to the scientific community. As an example how that can be done, Wylie introduces the case study of Kwaday Dän Ts'inchi, or the Long Ago Person Found,

emphasizing the possibility of a mutually beneficial partnership between the scientific and extra-scientific community.

The case study of Kwaday Dän Ts'inchi, or the Long Ago Person Found, goes back to 1999 when frozen remains of a young man were discovered in Northern BC, on the territory of CAFN (Champagne and Aishihk First Nations). This discovery presented an opportunity to renew pre-existing cooperation between CAFN and archaeologists together with provincial officials. The cooperation proved to be mutually beneficial, as it addressed the interest of CAFN to discover potential descendants while paying full respect to their local values of deceased. As their interests were acknowledged, they were willing to give consent to destructive testing and DNA analyses that were important for the archaeological research. The study discovered that young man lived between 1670-1850, he was 18-20 years old, he started his 100 kilometers long trip from the cost to the inland 3 days before he died and had several relatives in Wolf Clan.

The most interesting discovery of this joint effort was scientific confirmation of the first nations' oral tradition claims that coastal and inland community were connected with same clan affiliation identity. This finding, on the other side, called into question the assumption in archaeology that relates tribal identity to geographical locations, based on the Euro-American traditions. From this example, Wylie draws the conclusions on the benefits of diverse perspective that offer criticism from an external point of view and challenge similar views of the scientific community taken for granted. She identifies three epistemic benefits of collaborative practice from this example. First, it broadens the scope as to include questions from CAFN about possible descendants. Secondly, more resources are used, including CAFN background knowledge about local ecology and their oral history. And thirdly, the presumptions of geographical relation to tribal identity was re-examined and evidence that can be used was reconsidered, such as oral history in this case.

The extended transformative criticism that Wylie introduces, challenges framework assumptions but also gives unique insight and critical perspective from the standpoint of social margins. Wylie's example shows that extended external collaboration takes into account moral concerns, but also recognizes epistemic improvement as a result of critical external collaboration. Therefore, Wylie argues for reframing this fourth norm of tempered equality of epistemic authority to extend it to the criticism external to the scientific community. The extended version of transformative criticism that Wylie introduces challenges both framework assumptions and

gives unique insight and critical perspective from the different standpoint that includes moral concerns as well as improvement in epistemic practices.

5.2 Principle of democratic accountability

In order to examine if the principle of democratic accountability is attainable in the model of trustworthy scientific institutions model, our intention is to determine the role that scientific experts and policy makers have in policy decisions. The principle of democratic accountability primarily be examined in the light of inductive risk argument but within its limited scope of employing scientific expertise only to science advising (Douglas, 2009; Steele, 2012; Elliott and Steel, 2017). In addressing concerns raised by Betz (Betz, 2013; Elliott and Steel, 2017) as a proponent of value-free ideal, our aim is to show that democratic accountability is not endangered when scientific experts use value judgement but is an integral part of trustworthy institutions model.

One of the challenges to democratic accountability in the inductive risk argument is an unjustified power of scientific experts who use value judgments in informing policy while instead only democratically elected officials should make policy decisions (Steele, 2012; Betz, 2013). Betz's (Betz, 2013; Elliott and Steel, 2017) main claim in this context is that scientific experts in their advisory role to the policymakers do not have to rely on non-epistemic values nor are they morally required doing so.

The role of scientific experts should not extend above information providers because otherwise, it would jeopardize democratic accountability of science to the public. Only democratically elected officials are entitled to make policy decisions based on their value judgments. Therefore, Betz proposes that scientists in their role as science advisors should be explicit about uncertainties that are inherent in scientific findings in the form of "hedged" claims that themselves stay certain. It is not up to scientific experts to manage the inductive risk and make decisions in face of uncertainties. By claiming that non-epistemic values are neither necessary nor morally required his intention is to defend the value-free ideal of science.

While we would not dispute that policymakers as democratically elected officials should make policy decisions, we would propose a more beneficial role of science advisor as a broker

(Pielke, 2007) who presents various options and their impacts to the policymakers. In this process acknowledging uncertainties inherent in scientific findings should also emphasize scientific humility. Scientific advice is at the end only one of the input to the final policy decision among other relevant factors including taking risks in the face of uncertainties. However, Betz presumptions that scientific process does not entail non-epistemic values and that even presenting various options is completely value-free is misleading. Furthermore, as a consequence, it might have even more problematic stealth advocacy that masks values in neutrality and prevents openness about values.

Although there are several lines of arguments and directions in addressing these concerns we will mainly focus on the aspect of value-laden science expert advice within the democratic accountability principle. We claim that expert advice might be achieved in line with democratic accountability principle in three main ways. First, within the institutional trustworthy advising structures within which experts operate, second by ensuring openness about value judgments and third by including public interest in the process.

First, an example of science advice mechanisms that Carrier (Carrier, 2010) presents or the Science Advice Mechanism advising European Commission (SAM) have detailed and transparent regulations governing the science advice process established through democratically accepted practices. Furthermore, policymakers in the final instance do not have to accept science advice on specific policy issues. Both those instances indicate that the roles of individual experts values as articulated in the Betz argument are overstated.

Two additional aspects of transparency and call for public participation further bolster democratic accountability. Douglas (Douglas, 2009) in this regard recognizes that complete transparency and openness about value judgements in practical terms is not achievable, and suggests democratic accountability to be further bolstered by increased public involvement in research direction during the scientific process and in policy by framing the value positions. Community - based participatory approach has been employed in many social science disciplines together with consensus conferences, citizen engagement and other public participatory approaches that Douglas emphasizes and calls for their refinement. Further articulation of this essential part of public interest science will be examined in the following section within the broader context of inductive risk argument.

5.2.1 Inductive risk argument in the context of democratic accountability

Our approach to the discussion about inductive risk argument will be primarily focused on the scientific expertise relevant for policy advice and it will be examined in relation to the prospect of democratic accountability. Instead of employing inductive risk argument in the whole scientific process, as elaborated by Heather Douglas (Douglas, 2009, chapter 3), we tend to apply the discussion to the limited range within scientist's role as policy advisor (Steele, 2012; Betz, 2017).

In order to relate the discussion on the inductive argument in the context of democratic accountability, we will first focus on the formulation of the inductive risk argument. As indicated in the introductory remarks about the values in science, inductive risk argument is one of the possible places for the positive role of value in science. The initial argument (Rudner, 1953; Churchman, 1948) was further elaborated and extended by Heather Douglas (Douglas, 2000, 2009, 2017) from the theory acceptance to the whole "internal process" of science: the choice of methodology, gathering and characterization of data and interpretation of results.

The basic idea behind the inductive risk argument is that risk of error is present in all those stages, as there is never a complete certainty for correctly accepting hypothesis based on the available evidence. As Douglas clarifies

> "The uncertainty endemic in science arises from science's inductive and ampliative nature, where the evidence for any particular claim is never complete and the power of scientific generalizations is that they go beyond the evidence available, either by extending their descriptions to cases yet unseen or by positing causal relationships and/or explanatory theories that say more than the extant evidence can. It is through these extensions beyond the evidence that science gains its explanatory and predictive capacities" (Douglas, 2017, p. 81).

The same implications of incomplete evidential support and scientific generalizations constitute the uncertain nature of science requiring and justifying the use of non-epistemic values in scientific reasoning.

Scientists have to make social and ethical value judgements on how much evidence is required before accepting hypothesis by considering possible societal consequences of false positives and false negatives in relation to public policy (Douglas, 2000, 2008, 2009, 2017). The

internal reason for it is that epistemic and cognitive values cannot provide nonarbitrary conventions for the standard of evidence. The external reason is that due to societal consequences that scientific reasoning has on policymaking, scientists are morally required to use non-epistemic value judgements, but in open and explicit way in order to comply with democratic accountability (Douglas, 2008; Elliott and Resnik, 2014).

Internal standards, among which Douglas includes epistemic and cognitive values, cannot determine standards of evidence "because to employ one blanket standard fails to recognize the complex range of sufficiency judgements in science, and because the authority of science in society requires a consideration of the social and ethical implications of erroneous judgement" (Douglas, 2017, p. 81). Therefore social and ethical values play an important role for scientific expertise in the inductive risk argument. According to Douglas (Douglas, 2017), the accepted hypothesis does not imply reaching uncontested truth, but on contrary by its openness to critical scrutiny, ensures that the science is our most reliable source of knowledge.

One of the contextualization of the inductive risk argument comes into forth within the context of scientific expertise in providing science advice. In this regard, we analyze Betz proposal of using a hedged hypothesis (Betz, 2013) as a way of acknowledging uncertainties and presenting different possibilities that are context related. Our analysis suggests that the main goal of introducing hedged hypothesis - namely to defend value-free ideal - does not seems to be accomplished. However, further examination of the proposal might offer an inventive tool for presenting new innovative policy options in the realm of value-laden trustworthy science.

Betz's main claim is that scientific experts in their advisory role to the policymakers do not have to rely on non-epistemic values and are not morally required to do so. Their role should not extend above information providers because that would jeopardies democratic accountability of science to the public. Only democratically elected officials are entitled to make policy decisions based on their value judgements. Therefore, he proposes that scientists in their role as science advisors be explicit about uncertainties that are inherent in scientific findings in the form of "hedged" claims that nevertheless stay certain. It is not up to scientists to manage the inductive risk and make decisions in face of uncertainties.

By claiming that non-epistemic values are neither necessary nor morally required his intention is to defend value-free ideal of science. However, a closer analysis of the ways he proposes to achieve the goal of value-free ideal indicates that the evidence provided is not

sufficient for proving that inductive risk argument is unacceptable in the context of policy advice. We will address these concerns within two lines of argument.

First, Betz proposes that science advisors should use comprehensive sensitivity analysis for presenting different alternative choices to policymakers instead of making their own decisions. The method is applied in different contexts and its benefits for the policymaking we do not dispute. However, using implications of this method for supporting the value-free ideal is not so obvious. Because the very process of presenting a set of alternatives already entails specific sort of judgement, meaning that the choice of options could not be entirely free of all value judgements. Betz proposal of comprehensive sensitivity analysis entails specific choice of options in presenting different alternative choices to policymakers and therefore is not entirely free of all expert value judgements.

Robert Pielke (2007) also acknowledges that value judgements could not be avoided when science advisors present policy options. When he presents the role of "Honest Broker of Policy Alternatives" as one of four possible roles for science advisors, he acknowledges the presence of value judgements:

> "It is important to recognise that the Honest Broker of Policy Alternatives is very much an ideal type. In practice, having a truly comprehensive set of options would be overwhelming if not paralyzing. And any restricted set of options will necessarily reflect some value judgements as to what is included and what is not" (Pielke, 2007, p. 142).

As a response to the practical necessity of restricting available options, Betz (Betz, 2017) suggests overcoming it by the division of cognitive labor. However, his final possible solution of partial sensitivity analysis, in fact, confirms our claim that judgement is necessary for presenting a minimum range of possibilities among which policymakers would decide. Therefore, we claim that presenting different options, available solutions and entailed uncertainties that Betz suggests does not seem to imply that the very claims about those options and uncertainties are certain and value-free.

The second objection that could be directed to Betz probabilistic hypothesis has already been envisaged by Rudner (Rudner, 1953) who noted that assigning probabilities is also value-laden. Rudner already anticipated that critics could claim that scientists should not make value decisions in accepting the hypothesis, but only to assign probabilities and present them to

policymakers. Therefore, in the early formulation of his argument, he emphasized that value-laden decisions on sufficiency are also inherent in this process of assigning the probabilities.

However, acknowledging value-laden expert judgements should not hinder democratic accountability. Trustworthy scientific expertise might be better achieved by openly recognizing the existence of value judgements rather than acting in stealth advocacy disguising values under objectivity. Therefore, we should formulate the notion about the unjustified power of scientists who use their value judgements for informing policy instead of only democratically elected officials making policy decisions (Steele, 2012; Betz, 2013).

Following the above analysis, it seems that unjustified power of scientists could be of a greater risk if their value judgements are not acknowledged and if scientists are not open about their potential financial biases and conflicts of interest, rather than pretending that no value judgements are at place. This process of informing policy might then better serve democratically elected officials in making policy decisions. We will argue that democratic accountability could be achieved in three main ways within the institutional setting that ensures openness and public participation.

5.2.2 Openness about value judgements

We will argue that being open about value judgements that scientists make can only contribute to democratic accountability in policymaking processes. This is also one the aspect of inductive risk argument in the context of democratic accountability through openness about values (Douglas, 2009) that are inherent in the scientific process. Scientists do make value judgements, and acting as if their values are invisible will not make them disappear. It might instead mask certain interests under the veil of value neutrality enabling commercial interests to obstruct undesired regulations (Oreskes and Conway, 2010; Elliott and Resnik, 2014).

Along those lines, Douglas (Douglas, 2009) requires also from scientists to be transparent and explicit about their value judgements as an integral part of democratic accountability in science for policy.

> "In addition to keeping values to the indirect role in risk assessment, scientists should strive to make judgements, and the values on which they depend, explicit. Because policymaking is part of our democratic governance, scientific judgements made with these

> considerations have to be open to assessment by decisionmakers. Scientists should be clear about why they make the judgements they do, why they find evidence sufficiently convincing or not, and whether the reasons are based on perceived flaws in the evidence or concerns about consequences of error. Only with such explicitness can the ultimate decisionmakers make their decisions based on scientific advice with full awareness and the full burden of responsibility for their office" (Douglas, 2009, p. 155).

Openly presenting value judgements should enable public and policymakers to decide if they share the same views with the way scientist is waging the risks (Douglas, 2015) and also possibly subject them to critique from various perspectives, along the lines of transformative criticism that Longino (1990, 2002) introduces.

However, this account could potentially be criticized on several accounts, is it realistic to expect from each individual scientists to be aware and open about his value judgements throughout the research process. Secondly, tracing the values that are behind scientists claims and that should help the public to decide if they agree with them seems very complex procedure and very difficult for a non-expert to navigate. The unintelligibility of the procedure can hardly contribute to the achievement of public trust. De Melo-Martin and Intemann criticize this approach questioning the attainability of the proposed process where scientists would be fully aware of their values throughout the scientific process (de Melo-Martín and Intemann, 2016). They argue that openness about value judgements assumes that scientists themselves are aware of their value judgements, which might not be the case, specifically if value judgements are at place in the whole research process.

Furthermore, at the point when scientific results influence policymaking to "backtrack" their value decisions throughout the research process or even assess possible outcomes if different value judgements were employed does not seem plausible. It would require specific expertise but even then unpredictability of conclusions that other value judgements might yield would pose a challenge itself. Even if backtracking of value judgements is possible, what if at the end it results in a disagreement about those value judgements that have been made in the research process. Should policymakers then disregard the research results and make uninformed policy decisions or should they proceed according to research results although they are not in line with democratic values. Further proposal of involving public throughout the research process implies that scientists would execute public value judgements thus neither giving a convincing account

against value-free ideal (de Melo-Martín and Intemann, 2016). In this regard, Douglas (Douglas, 2009) also recognizes that complete transparency and openness about value judgements in practical terms is not achievable.

Openness about value judgements seems to pose a specific challenge and we will propose that the aspect of openness would be better addressed within institutional structural framework than only by focusing on the individual scientist. Examples on the endocrine disrupting chemicals (Elliott and Resnik, 2014) or tobacco industry case study (Proctor, 1995, 1996, 1999, 2008, 2011) clearly illustrates how the blurring conflict of interests and financial biases behind the value-free disguise can have serious implications for policy regulations and consequently for public trust in science. Next discussion on high standards of evidence will bring us to the level of structural institutional settings needed for adequately addressing the principle of openness in scientific endeavor.

5.2.3 Institutional setting and high standards

The examples on the endocrine disrupting chemicals (Elliott and Resnik, 2014) have demonstrated how setting high standards of evidence and demanding ever more evidence can influence regulation and consequently result in harmful effects for the public. In this specific case, setting high standards for evidence before regulations can be imposed was closely related to financial biases. Cases like that or tobacco industry example certainly have a great impact on public mistrust in science.

The issue against high standards of evidence is further elaborated by Steel in terms of second-order uncertainty. He specifies the second-order uncertainty as "uncertainty about an assessment of uncertainty" (Steel, 2016, p. 697) critically addresses the objections raised within the concept of inductive risk argument (Betz, 2013; Morrison, 2014: Parker, 2014). In his generalization of the inductive risk argument Steel (2016) argues against high standards of evidence as scientific uncertainty might suspend judgement and related policy action and thus hinder the problem at hand to be addressed. That is particularly important when the policy implications have an immense impact on the public.

Steel further extends inductive risk argument to its structural version (Steel, 2016, p. 697) by considering also together structural and community aspects of inductive risk (Wilholt, 2009,

2013: John, 2015; Steel, 2016) that can serve as a good starting point for the trustworthy institutional design. Based on Steel's structural argument we further suggest institutional design to be flexible and open to transformation and institutional learning based on the possible critical aspects and failures. The institutional mechanisms in place should ensure institutional trustworthiness and enable well-placed trust within the scientific community and on behalf of the public.

Furthermore, the discussion about inductive risk argument should consider its implications for the policy as situated in the practical context of applications. Rarely will scientific results published in a paper have a direct influence on the decision making process. The science advisory processes are rather implemented in various institutional settings that often not take into consideration scientific results from various disciplines related to the challenge at hand. Moreover, within science advice mechanisms scientists have a different role as science advisors and are primarily using their scientific expertise to tackle the specific practical problem. Carrier and Krohn (2018) analysis of particular case study of the practical work of one such scientific committee gives more instructive perspective on the ways in which scientific results are being used.

The process of devising scientific advice consults scientific results from various disciplines that can then be interpolated, since they might be partially relevant to the practical problem at hand and can also lead to different, sometimes even disparate solutions. As the study shows (Carrier and Krohn, 2018) the commission composed of scientific experts can also produce new models that will be useful for the specific model instead of using generalized knowledge.

Overestimating the impact that an individual scientist and his values have on the policy process fails to consider two main aspects. First, that individual scientist is a part of the scientific community and operates within methodological standards of specific disciplines. Second, that in his role as science advisor, he operates again mainly within a specific system or institutional framework. Often he works together with other stakeholders and exercises additional scientific expertise resulting with additional, context-specific knowledge production. Therefore, we argue that when considering inductive risk argument the role of scientists would be more adequately addressed in its institutional context. That at the same time does not exclude or suspend his value judgements that might be crucial for the healthy critical transformative criticism within dynamically operating trustworthy institutional scientific design.

5.2.4 Publics

The third aspect within the discussion on democratic accountability perspective is the role of the public and public involvement on topics that directly impact them, be it in the early process of setting the research agenda or later in the conducting the research. Along those lines Douglas (Douglas, 2009 chapter 8) proposes increased public involvement in research direction, helping scientists during the scientific process and in policy by framing the value positions as a way of increasing democratic accountability. Introducing this aspect of public involvement might give a better prospect for democratic accountability since openness and transparency about scientists' value judgements cannot be enough. When suggesting this aspect, Douglas also recognizes that complete transparency about value judgements in practical terms is not achievable, or that it would even be useful in its extensive form.

Being accountable to the public is one more aspect that Douglas introduces into relation between scientific practices and public involvement

> "As scientists come to understand the full implications of their increasingly public role, they need to be open to genuine public input on how to best fulfill that role. One cannot claim to be publicly accountable but unmoved by public criticism. Nor can one claim autonomy from public accountability if one wields power in the public realm. In short, just as scientists want the public to listen to them, they will, in turn, have to learn to listen to the public" (Douglas, 2009, p. 172).

Although, as previously elaborated, the prospect of placing discussion of public involvement in the context of institutions have a better prospect than focusing on the relation solely to individual scientists, specific aspects of the analytic-deliberative process will be applicable to both approaches.

Therefore, we will further analyze the modalities and the prospect of public involvement in the context of democratic accountability based on the formulation of both theoretical account and practical application of the analytic-deliberative process (Douglas, 2009, chapter 8). Concludingly, we will examine practical attainability of this analytic-deliberative approach in the Valdez case study.

A theoretical account of the analytic-deliberative process in risk assessment can best be explained by focusing on the differentiation between analytic and deliberative part of the process

as elaborated by Douglas (2009, chapter 8). In the analytic-deliberative process, analytical part and deliberation mutually influence and build on each other. Analytic part of the process performs neutral and objective analysis that is in accordance with standards of a particular field. On the other side, deliberation is interactive and draws from different perspectives. Part of deliberative processes is also to decide which analysis will be utilized.

Deliberation processes enable on-going input from the public, or it can involve different groups of experts such as in peer review process. Deliberation can also take place in mixed groups between experts and non-experts, or more precisely with the public that will be affected by the final decision coming from the analytic-deliberative process.

> "The public nature of expert judgements, accountable to all interested parties, keeps them from becoming sources of contention or accusation later on. This can greatly bolster public trust in risk assessments. In addition, making such judgements clear is not a threat to scientific integrity as long as the role of values is kept to their legitimate forms, not suppressing or inventing evidence. Then the appropriate norms for values in scientific reasoning are upheld" (Douglas, 2009, p. 163).

This approach can potentially be beneficial in addressing public views and public participation in the scientific and policy processes that will have direct impacts on people's lives. Therefore, we are interested what are potential benefits and downsides of applying this approach instead of numerous other approaches on public involvement and what implications it will have in the context of trustworthy scientific institutions.

At first, there are several valuable aspects of the prospect of democratic accountability when using this analytic-deliberative processes. First, involving the publics within deliberative processes reinforces the democratic accountability of expert judgement since it is clearly elaborated throughout the deliberation process. It therefore, contributes to building trust from the very beginning of the process and ensures greater acceptance of the final decision because it has been achieved in an inclusive way. Secondly, scientific integrity is preserved since deliberation does not determine the results of expert analyses. And finally, analytic-deliberative processes have only advisory role, public officials are still the one who will make the final decision. Therefore, democratic processes are also not jeopardized.

Moreover, in terms of its application, this approach does not have to be directed only to individual scientists but can be taken on board also in a more structured approach. Douglas

(Douglas, 2009, chapter 8), along those lines, also examines scientific input within specific analytic-deliberative processes of risk assessment or consensus conferences that are structurally embedded and not only focused on individual scientists.

When it comes to practical implementation, however, this proposal might encounter several difficulties. One of the possible critics is that the diversity of advisory boards might be difficult to achieve either because relevant expertise is not available or because economic or political factors can prevent affected parties to actually come on board and engage (de Melo-Martín and Intemann, 2016). De Melo-Martin and Intemann illustrate it in the case of IPCC (Intergovernmental Panel on Climate Change) that is often being criticized as being composed of white male scientists.

Another critic that relates to the example of IPCC maintains that this way of public engagement might be possible at the local level, but could be problematic for the decisions that have to be made on the global level such as in IPCC. In issues where policy decisions would have implications on broader public and it would not be feasible to achieve closer involvement of public with scientists, Douglas (Douglas, 2009) therefore proposes to consider other options.

Some of those possible options would be participation approach or consensus conferences (Douglas, 2005). Participatory risk assessments could be conducted at the local level and then fed it into the national level. Several other solutions have already been developed by social scientists, such as consensus conferences first initiated in Denmark. They engage randomly chosen citizens as representatives of the population and educate them on specific policy issue at hand, who then ask experts the questions and raise their concerns. At the final stage, they develop their consensus statements on the value trade-offs. The results of consensus conferences are then used as an instructive part for the research regarding the value judgements.

Potential of analytic-deliberative processes and various modes of public involvement have been recognized since

> "Indeed, there are many examples in which stakeholder input has been successfully incorporated in establishing and refining aims of research in a way that is then used to justify particular methodological decisions (Shrader- Frechette, 2007). As Douglas herself notes, community-based participatory research has been used in many social science disciplines, where researchers work with community advisory boards composed of

representatives of groups affected by the research" (de Melo-Martín and Intemann, 2016, p. 517).

They further refer to already prominent examples of community-based participatory research in several fields from health research on HIV to climate change research:

> "This has been common, for example, in both national and global research on HIV/AIDS prevention and treatment. In these cases, advisory boards participate not merely in crafting policy recommendations for, for example, needle-exchange programs or HIV education programs. Rather, they play a role at various stages throughout the research process: in formulating what the policy aims and the priorities of the research should be, giving feedback on the extent to which methodological decisions sufficiently advance those aims (such as clinical trial methodology), and providing critical feedback on assumptions that scientists have made in interpreting data (Epstein 1996). Similarly, in the context of climate change research, there are increasing efforts to incorporate stakeholder input throughout the research process (Kloprogge and Van Der Sluijs 2006; Tang and Dessai 2012; Kirchhoff, Lemos, and Dessai 2013). The UK Climate Impacts Programme, for instance, has developed mechanisms for working with stakeholders to identify adaptation needs and receive critical feedback on modeling strategies to produce more "useable knowledge" (Tang and Dessai 2012)." (de Melo-Martín and Intemann, 2016, p. 517).

There are numerous benefits of community-based participatory research and consensus conferences, but also Douglas (Douglas, 2009) recognizes their further critical points and calls for their refinement. There are indeed numerous further implications of participatory approaches and involvement of publics, and one of them is their potential use within the model of trustworthy institutions. In order to illustrate attainability of participatory approaches on the institutional level, we will illustrate it with the Valdez case study (Douglas, 2009).

By focusing on this case study our aim is to show how trustworthiness could be achieved at the institutional level when particular properties of institutional design are at place making trustworthy relations between various actors possible. The Valdez case study triggered our interest because it shows how trustworthy relations between various stakeholders could be achieved illustrating it on a very simple scale, in regards to what kind of tug vessels should be

used. Analytic-deliberative process, in this case, achieves more than satisfactory results that are beneficial for all involved parties.

First significant point is that all parties were focused on the challenge at hand and were determined to find a solution on what kind of tug vessels should be deployed and they all financially contributed to the study. Among participants were also policymakers who were vigilant to adopt the final results. On the side of the risk assessment team, scientists from both industry and citizens' organization were part of the team and jointly conducted the risk analysis. The steering committee, on the other hand, was assembled of diverse representatives of the public, policymakers and industry and they met regularly to direct the risk assessment along the analytic-deliberative process (Douglas, 2009, p. 164-166).

In this case study, analytic-deliberative process that was at place showed to have achieved policy decision in a trustworthy way. Neither unsound science, conflict of interest, macro biases or public mistrust were an issue. Risk assessment results were not presented at the very end of the scientific process or with possible predetermined results that could stem from the possibly hidden conflicts of interest. On contrary, public contribution together with contribution from other stakeholders from industry was present from the very beginning of the analytic-deliberative process thus ensuring genuine responsiveness and responsibility in the co-production practice. Trustworthy process ensured that specific interests do not predetermine the final results, preserving scientific integrity and public participation.

This case study presents ideal circumstances where all parties are determined to find a solution, that surely cannot be easily applied to all other cases. But our intention is not to present the case that will have universal application. Here we encounter the eternal quest on how to make the context general as possible in order to be relevant for various cases and on the other side as specific as possible in order to be of real practical use. The endeavor of satisfying both conditions at the same time is almost impossible and it is surely not our intention. The scope of this research proposal is limited to the outlining the basic structural ingredients that would make a good base for the prospect of trustworthy institutions without pretentious of applying it to each specific particular case where further specifications and refinements would be necessary.

5.3 Conclusion

The trust and trusting relations are the precious glue for society and a social capital. As Confucius would advise his disciple Tzu-Kung "Abandon weapons first, then food. But never abandon trust. People cannot get on without trust. Trust is more important than life." In addressing the challenge of trust, we reframe the problem from restoring or improving trust to the question about trustworthiness. On that ground then we defend the thesis that the trustworthy by design model is better suited to address the question of distrust and misplaced trust.

This reverse approach based on trustworthiness is not new, several authors acknowledged its importance in interpersonal relations (Baier, 1986; Pettit, 1995; Jones, 1996, 2012; McGeer, 2008), as well as from institutional and social perspective (Hardin, 1996, 2002; O'Neill, 2002a, 2013, 2014; Potter, 2002). Furthermore, the conceptual framework on the difference between concepts of trust and trustworthiness we based on the work of Onora O'Neill (2002a, 2002b, 2013, 2014) emphasizing that we can place trust well only if we trust the trustworthy (O'Neill, 2013). Therefore, we argue for giving priority to the concept of trustworthiness over trust because it has distinguished properties that could be reflected upon enhancing the prospect of placing trust wisely.

The argument that we present in favor of the trustworthy by design model, differs from prominent approaches as it suggests redefining the main properties of the concept of trustworthiness so that it accommodates more inclusive aspects of responsibility and responsiveness to the diversity of perspectives. Moreover, our intention is to test the model applying it to two case studies on algorithmic systems and scientific institutions to demonstrate its significance in the nexus between science, society and policy and within the 4th industrial revolution. Proposed redesigning of the systems and institutions should enable two-sidedness of trusting between trustee and trustor, publics and systems based on diversity and inclusiveness.

In the first chapter, we directly apply the trustworthy by design model to the algorithmic systems to stress the challenge of trustworthiness that is even more profound in the new forms of trust in digital platforms than in previous forms of trust in individuals and in institutions. Due to the unbearable easiness of trusting in digital platforms, our intuitive trusting mode is even harder to challenge, so we pause to inquire about the trustworthiness of the algorithmic systems that run in the background.

Next to being easy to place our trust in digital platforms it also at first seems like a better prospect of placing our trust wisely. Judging "the other person's competence and honesty, as well as their reliability, in the relevant matter" (O'Neill, 2013, p. 238) seems only to become easier with digital platforms through consulting rating and reviews. However, a closer analysis of the algorithmic systems that run in the background uncovers more complex issues since algorithms are not neutral and since it is not clear who should be held responsible in the complex interrelation between human and artificial agents.

Our mapping exercise of the indicators of changing trust in the context of algorithmic systems focuses primarily on the decision-making algorithms such as personalization algorithms (Hildebrandt, 2008; Newell and Marabelli, 2015; Taddeo and Floridi, 2015), profiling algorithms (Hildebrandt, 2008), machine learning algorithms, (Tutt, 2016; Burrell, 2016), negotiation algorithms (Raymond, 2015) to clinical decisions algorithms in computer-based diagnostic systems (Mazoue, 1990). Next to obvious benefits in terms of efficiency and enhanced capabilities, we identified two main challenges in algorithmic systems.

First is related to values in the design of algorithms, technical constraints or values in the process of their implementation. Far from being neutral, algorithms are value-laden. Values behind algorithms are identified in the form of racial biases, where search algorithms are discriminating people of color (Noble, 2018), racial biases in predictive policing (Lum and Isaac, 2016, Ferguson, 2017, Angwin et al., 2016), in clinical trials (Kurt et al., 2016), etc. Gender biases are present in adds that are targeting man more than female for higher paid jobs (Datta et al., 2015, Campolo et al., 2017), or careers in STEM (Lambrecht and Tucker, 2016). Also word embedding that is often used in machine learning algorithms for ranking in the Web search (Nalisnick et al., 2016) or CV analysis (Hansen et al., 2015), can exhibit gender biases (Bolukbasi et al., 2016) such as found in *man* appearing more often next to *computer programmer* and *woman* next to *homemaker.* Based on the analysis of the value-laden algorithms we identify indicators relevant for changes in public trust, such as biases, discrimination, fairness, diversity, and, inclusiveness.

Analysis of the second challenge on responsibility in distributed systems reveals indicators of changes in public trust such as governing challenges, regulation, accountability, transparency and public engagement. Moreover, the concept of individual responsibility loses its relevance for the complex systems of human and artificial agents, where the concept of

distributed agency (Floridi 2013; Floridi and Taddeo, 2016) is better suited to hold "all agents of a distributed system, such as a company, responsible. This is key when considering the case of AI, because it distributes moral responsibility among designers, regulators, and users. In doing so, the model plays a central role in preventing evil and fostering good, because it nudges all involved agents to adopt responsible behaviors" (Taddeo and Floridi, 2018, p. 751).

Based on the mapped indicators of changing trust in algorithmic systems, we argue that algorithms can neither be trusted or trustworthy. Although algorithms - specifically machine learning algorithms that are partially autonomous - could be held accountable, they do not possess intentionality that is unique only to human moral agents. We argue, that instead, trustworthiness should be directed to the complex interrelation between human and artificial agents. Based on that background, the trustworthy by design model refers to algorithmic systems by encoding distributed responsibility and responsiveness directly in the early process of algorithmic design in the interplay between algorithms and humans. Blockchain technology serves as a case study of the potential practical implementation of the participatory and inclusive practices through encoding responsiveness and responsibility in the design of algorithms that might have a big impact for the prospect of trustworthiness and trusting relations.

In the second chapter, we then draw back to close critical analysis of the current state of the art on approaches on trust, covering a review and analysis of the extant literature in the research on trust, various philosophical dimensions of trust, and the difference between trust and trustworthiness. Apart from several unifying characteristics, the accounts of trust diverge significantly in determining the motivational causes of trust, from goodwill (Baier, 1986; Jones, 1996), moral commitments (McLeod, 2002; Nikel, 2007; Cohen and Dienhart, 2013), interests (Hardin, 2002), social norms and constraints (Dasgupta, 1988; Hardin, 2002; O'Neill, 2002a) to motivational neutrality (Jones, 2012). Following through different motivational aspects as well as through metamorphoses of trust from trusting relations between individuals, to forms of trust in institutions and transformations of trust in digital technologies, we defend three hypothesis.

First, analysis of the goodwill account of trust (Baier, 1986, 1991, 2013; Jones, 1996, 2013, 2017) points to the significant role that *strong thin* mode of trust (Hosking, 2014) could play in understanding transformations of trust. In order to determine if trust is emotional or cognitive, we first start with Jones's (1996, 2013) affective account of trust that consists of both cognitive aspect of trust as *expectation* that the trustee will be favorably moved by the realization

that we are counting on her and *affective attitude* of optimism about trustee's goodwill and competence, the aspect of trust that Jones gives priority to.

We then conduct analysis of influence of pre-existing beliefs and emotions on public trust in science (Golenberg, 2016), including the evidence from experimental psychology (Kahneman, 2013), insights on how beliefs in scientific claims do not relate so much to expertise but to trust in the source of information and relation to our beliefs (Kahan, 2013, 2017) or evidence from cognitive science (Sloman and Fernbach, 2017) on knowledge illusion and protecting our communities of similar beliefs and attitudes. Based on that analysis we argue for abandoning the knowledge deficit hypothesis and for acknowledging significant role of both affective, emotional as well as cognitive aspect in the context of trust.

The final analysis of the encapsulated interest view introduced by Hardin (2002, 2006) challenged the attainability of interest being the main motivation in the context of institutions. Based on the identified shortcomings and critical points of different concepts of trust, instead of abandoning trust in institutions, we formulate and elaborate in more detail the new trustworthy by design model. We argue that it should be applicable both in the institutional setting as in the context of digital technologies that was applied in the first chapter.

Based on the identified shortcomings of the various concepts of trust, we propose properties in redesigned trustworthy institutions that form the trustworthy by design model. The model addresses the main challenge of placing our trust blindly or misplacing trust in the untrustworthy that can result in disappointment, misuse or even exploitations. In our articulation of the trustworthy by design model in the context of institutions we first focus on the priority of trustworthiness based on normative and empirical aspects and intentionality.

We take intentionality to be prerequisite for placing our trust well based on O'Neill's approach who "rejects an "attitudinal" account of trust, since she takes trust something that is "intentionally" placed in the trusted, where the placing may or may not be intelligently done" (Baier, 2013, p. 181). In discussing basic properties of institutional design that would ensure trustworthiness, next to aspects of competence, honesty, and reliability (O'Neill, 2013) we propose responsibility in distributed agency and responsiveness as two features that enable participatory practices. Further reconceptualization of the critical aspects of the mechanisms of accountability and transparency should enable the mechanisms to be used as a means of judging where to place trust and enabling intelligibility of complex institutions.

In the final instance, our analysis leads to developing an alternative trustworthy by design model in the context of institutions. We propose a model as a three-part relation in which B is trusted by A for specific thing X. Where B as an institution signals its trustworthiness by encoding the responsiveness and responsibility in the uptake of concerns and values that A has regarding specific thing X. Trustworthy by model design formulated as such should be better suited to address the identified challenges of distrust and misplaced trust in the context of algorithmic systems and scientific institutions.

In the last two chapters, we apply the trustworthy by design model to the context of scientific institutions where we model our concept of trustworthy scientific institutions. We place the discussion within the broader framework of values in science in the interplay between science, society, and policy. Research controversies make it evident that together with benefits there are possible risks for society implied in epistemic dependence in science. Due to the lack of capacity to verify each of scientific claims (such as the earth is round) we often have to rely for our knowledge on the testimony of scientific experts and others (Hardwig, 1991). Experts might be in disagreement and scientific claims contested. Confirmation biases, potential unwarranted consequences of politicization and commercialization of science all might call in question trust in science and be possible causes of mistrust among the public.

In the chapter four we compare two ways of addressing this challenge: (1) trust in individual experts based on the account of individual expert responsibility (Douglas, 2008, 2015) and (2) institutional distrust approach - based on the institutionally organized skepticism (Merton, 1938; Bouchard, 2016) and transformative criticism (Longino, 1990, 2002). Our line of argument, however, departs from the prospect of promoting trust in individual expertise because it might fail to address the question of how to place trust wisely, while institutional distrust approach does not sufficiently acknowledge the conditions of the current societal context in which science takes place.

The analysis of those two approaches we interrelate with the tobacco industry case study as a means of identifying indicators that are influencing changes of public trust in science. Based on our mapping exercise we identify main indicators of changing trust in science in framing the problem and the scientific agenda, neglecting diverse voices, confirmation, and macro-biases and using uncertain nature of science for delaying regulations with potentially harmful consequences for the publics.

In the fifth chapter, we suggest implementing the trustworthy by design model to scientific institutions as a way of addressing indicators of changing publics' trust in science related to the broader societal context of science. The proposed model focuses not only on "internal" aspects of conducting science but emphasizes the significance of its "external" and organizational structure of its scientific institutions where the real focus on changing public trust in science should be placed. The introduced concept is intended to cover a broad array of institutional setting in which scientists operate from academia, advisory structures to the industry.

Through extra-scientific critical contributions (Wylie, 2014) as an extension of the transformative criticism (Longino, 1990, 2002) it should facilitate trustworthiness by encoding the principles of responsibility and responsiveness into redesigning scientific institutions. At the same time based on the analysis of the inductive risk argument (Douglas, 2000, 2009), it should ensure democratic accountability in the institutional context through openness and public engagement. Our analysis uncovers various forms of underlying ignorance, either as exclusion or active production of doubt raising several questions related to the objectivity of science. By applying the trustworthy by design model we argue that trustworthy science institutions require reformulation of the notion of objectivity in science to ensure responsibility, responsiveness and democratic accountability.

After devising the trustworthy by design model our intention in applying it in the two case studies in the algorithmic systems and the context of scientific institutions was at least to shed a light on the possible prospects of their redesigning capacities in the context of trustworthiness. However, we still did not overcome the challenge of devising the model as general as possible to be applicable to a variety of cases and specific as possible to have value in practical use. The further research still lies ahead in refining the constitutional properties of the suggested model as well as testing it in the broader number of case studies both in the two suggested fields in digital technologies and scientific context, as well as in other relevant areas. The hope is that the proposed thesis might be at least partially beneficial in this endeavor.

References

Anderson, M. and Anderson, S. L. (2014). Toward ethical intelligent autonomous healthcare agents: a case-supported principle-based behavior paradigm. Available at: http:// doc.gold.ac.uk/aisb50/AISB50-S17/AISB50-S17-Anderson-Paper.pdf

Angwin, J. Larson, J. Mattu, S. and Kirchner L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baier, A. C. (1986). Trust and Antitrust. *Ethics* (2). 231-260.

Baier, A. C. (1991). Trust and Its Vulnerabilities and Sustaining Trust. Paper presented at the *Tanner Lectures on Human Values,* Salt Lake City, 1991.

Baier, A. C. (2013). What is Trust? in *Reading Onora O'Neill*, ed. Archard, Deveaux, Manson, and Weinstock, London: Routledge: 175–85.

Barabas, C. Dinakar, K. Ito, J. Virza, M. Zittrain, J. (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. Available at: arXiv:1712.08238v2 [cs.LG]

Barker, C. (2015). Is blockchain the key to the Internet of Things? IBM and Samsung think it might just be. ZDNet [Online]. Available at: http://www.zdnet.com/article/is-blockchain-the-key-to-the-internet-of-things-ibm-and-samsung-think-it-might-just-be/

Barnet, B.A. (2009). Idiomedia: The rise of personalized, aggregated content. Continuum 23 (1): 93–99.

Barocas, S. and Selbst, A.D. (2016). Big Data's Disparate Impact. 104 *California Law Review* 671-732. Available at SSRN: https://ssrn.com/abstract=2477899 or http://dx.doi.org/10.2139/ssrn.2477899

Betz, G. (2013). In Defence of the Value Free Ideal. *European Journal for Philosophy of Science* 3:207–220.

Betz, G. (2017). Why the Argument from Inductive Risk Doesn't Justify Incorporating Non-Epistemic Values in Scientific Reasoning, in Current Controversies in Values in Science, eds. Kevin Elliot and Daniel Steel. New York: Routledge. 94-111.

Biddle, J. (2013). State of the Field: Transient Underdetermination and Values in Science. *Studies in History and Philosophy of Science*. 44: 124–133.

Blackburn, S. (1998). *Ruling Passion: A Theory of Practical Reasoning*. Oxford: Clarendon Press.

Bergman, A. Andersson, A. M. Becher, G. van den Berg, M. Blumberg, B. Bjerregaard, P. Bornehag, C. G. Bornman, R. Brandt, I. Brian, J. V. Casey, S. C. Fowler, P. A. Frouin, H. Giudice, L. C. Iguchi, T. Hass, U. Jobling, S. Juul, A. Kidd, K. A. Kortenkamp, A. Lind, M. Martin, O. V. Muir, D. Ochieng, R. Olea, N. Norrgren, L. Ropstad, E. Ross, P. S. Rudén, C. Scheringer, M. Skakkebaek, N. E. Söder, O. Sonnenschein, C. Soto, A. Swan, S. Toppari, J. Tyler, C. R. Vandenberg, L. N. Vinggaard, A. M. Wiberg, K. Zoeller, R. T. (2013). Science and policy on endocrine disrupters must not be mixed: a reply to a "common sense" intervention by toxicology journal editors. *Environ Health* 12:69; doi:10.1186/1476-069X-12-69.

Bloor, D. (1976). The Strong Programme in the Sociology of Knowledge. In D. Bloor, ed., *Knowledge and the Social Imagery*, London: Routledge and Kegan Paul, 1–19.

Bolukbasi, T. Chang, K-W. Zou, J. Saligrama, V. Kalai, A. (2016). Man is t Computer Programmer as Woman is to Homemaker? Bebiasing Word Embeddings. Available at: arXiv:1607.06520v1 [cs.CL]

Botsman, R. (2017). *Who Can You Trust?– How Technology Brought Us Together and Why It Might Drive Us Apart*, New York: PublicAffairs

Bouchard, F. (2016). The Roles of Institutional Trust and Distrust in Grounding Rational Deference to Scientific Expertise. *Perspectives on Science* 24(5), 582-608.

Brey, P. and Soraker, J. (2009). Philosophy of Computing and Information Technology. *Vol. 14 of the Handbook for Philosophy of Science.* Elsevier.

Brigandt, I. (2015) Social values influence the adequacy conditions of scientific theories: beyond inductive risk, *Canadian Journal of Philosophy*, 45:3, 326-356, DOI: 10.1080/00455091.2015.1079004

Buechner, J. and Tavani H. T. (2011). Trust and Multi-Agent Systems: Applying the "Diffuse, Default Model" of Trust to Experiments Involving Artificial Agents. *Ethics and Information Technology,* 1-13.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society* 3(1): 1–12.

Buterin, V. (2013). Ethereum: A Next-Generation Generalized Smart Contract and Decentralized Application Platform, *Blockchain Papers: Curated Cryptoasset Publications*, accessed September 30, 2018, https://blockchainpapers.org/items/show/2.

Campolo, A. Sanfilippo, M. Whittaker, M. Crawford, C. (2017). AI Now 2017 Report, AI Now Institute at New York University Available at: https://www.microsoft.com/en-us/research/uploads/prod/2018/02/AI_Now_2017_Report.pdf

Cardona, B. (2008). 'Healthy Ageing' policies and anti-ageing ideologies and practices: on the exercise of responsibility. *Medicine, Health Care and Philosophy.* 11(4): 475-483.

Carrier, M. (2010). Scientific Knowledge and Scientific Expertise: Epistemic and Social Conditions of Their Trustworthiness. *Analyse and Kritik.* 32(2): 195-212.

Carrier, M. (2016) Social Organization of Science in the Oxford Handbook of Philosophy of Science, Edited by Humphreys, P. DOI: 10.1093/oxfordhb/9780199368815.013.43

Carrier, M. and Krohn, W. (2018). Scientific Expertise: Epistemic and Social Standards – The Example of the German Radiation Protection Commission. *Topoi. An International Review of Philosophy*, 37: 55-66

Castelfranchi, C., and and Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model.* Chichester: John Wiley and Sons.

Churchman, C. W. (1948). *Theory of Experimental Inference.*New York: Macmillan.

Citron, D. K. and Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review,* Vol 89 (1)

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., and Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, 33(7), 1139-1147. DOI: 10.1377/hlthaff.2014.0048

Cohen, M. A. and J. Dienhart, (2013). Moral and Amoral Conceptions of Trust, with an Application to Organizational Ethics. *Journal of Business Ethics* 112: 1–13.

Coleman, J. S. (1990). *Foundations of Social Theory*. Boston: Harvard University Press.

Cook, K. R., Hardin, R. and Levi, M. (2005). *Cooperation Without Trust?* New York: Russell Sage Foundation.

Crawford, K. and Schultz J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55 (1): 93. Available at: https://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4

Crawford, K. (2016). Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology and Human Values* 41(1): 77–92.

Dasgupta, P. (1988). Trust as a Commodity, in *Trust. Making and Breaking Cooperative Relations*:49-72 (ed.) Gambetta, D. Oxford: Basil Blackwell.

Datta, A., Tschantz, M. C. and Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* (1): 92-112.

De Melo-Martín, I. and Intemann, K. (2016). The risk of using inductive risk to challenge the value-free ideal. *Philosophy of Science*. 83(4), 500-520. doi:10.1086/687259

Dietrich, D. R. von Aulock, S. Marquardt, H. Blaauboer, B. Dekant, W. Kehrer, J. Hengstler, J. Collier, A. Batta, G,G. Pelkonen, O. Lang, F. Nijkamp, F. P. Stemmer, K. Li, A. Savolainen, K. Wallace, H. A. Gooderham, N. Harvey, A.(2013). Scientifically unfounded precaution drives European Commission's recommendations on EDC regulation, while defying common sense, well-established science and risk assessment principles. Food Chem Toxicol. 62:A1–A4.

Dorato, M. (2004). Epistemic and Nonepistemic Values in Science. In Machamer P. and Wolters G. (Eds.), *Science Values and Objectivity* Pittsburgh, Pa: University of Pittsburgh Press: 52-77. doi:10.2307/j.ctt5vkg7t.7

Douglas, H. (2000). Inductive Risk and Values in Science, Philosophy of Science, 67(4): 559-579.

Douglas, H. (2005). Inserting the public into the science, in *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, Sociology of the Sciences, vol. 24, 156-169, Springer. Printed in the Netherlands.

Douglas, H. (2006). Norms for Values in Scientific Belief Acceptance, Available at: http://philsci-archive.pitt.edu/3024/

Douglas, H. (2008). The role of values in expert reasoning. *Public Affairs Quarterly*. 22(1), 1-18.

Douglas, H. (2009). Science, Policy, and the Value-Free Ideal, University of Pittsburgh Press.

Douglas, H. (2013). The Value of Cognitive Values, Philosophy of Science, 80:796–806

Douglas, H. (2015). Politics and Science: Untangling Values, Ideologies, and Reasons. *The Annals Of The American Academy Of Political And, Social Science,* 658296.

Douglas, H. (2016). Values in Science. In (Ed.), *The Oxford Handbook of Philosophy of Science*.: Oxford University Press,. Retrieved 1 Oct. 2018, from

http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199368815.001.0001/oxfordhb-9780199368815-e-28.
Douglas, H. (2017). Why Inductive Risk Requires Values in Science, in Current Controversies in Values in Science, eds. Kevin Elliot and Daniel Steel. New York: Routledge. 81-93.
Edelman. (2017). Edelman Trust Barometer. [online] Available at: https://www.edelman.com/trust2017/ [Accessed 28 Sep. 2018].
Edelman. (2018). 2018 Edelman Trust Barometer. [online] Available at: https://www.edelman.com/trust-barometer/ [Accessed 28 Sep. 2018].
Eurobarometer Special Survey (2010). *Science and Technology,* No 340. [online] pp.1-19. Available at: http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_340_en.pdf [Accessed 15 Sep. 2017].
EC (European Commission). (2013). Commission Recommendation of XXXX: Defining Criteria for Endocrine Disruptors.
European Commission (2017). *Standard Eurobarometer 87*. Eurobarometer Surveys. [online] Available at: http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/STANDARD/surveyKy/2142 [Accessed 15 Sep. 2017].
Elliott, K. C. and Resnik, D. B. (2014). Science, Policy, and the Transparency of Values. *Environmental Health Perspectives* 122:647–650; http://dx.doi.org/10.1289/ehp.1408107
Elliott, K. C. and Steel, D. (eds.) (2017). *Contemporary Controversies in Values and Science.* New York: Routledge.
Eyal, I. and and Sirer, E. G. (2013). Majority is not Enough: Bitcoin Mining is Vulnerable. Available at: https://arxiv.org/abs/1311.0243
Ferguson, A.G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press.
Feyerabend, P. (1962). Explanation, Reduction and Empiricism, in H. Feigl and G. Maxwell (ed.), *Scientific Explanation, Space, and Time*, (Minnesota Studies in the Philosophy of Science, Volume III), Minneapolis: University of Minneapolis Press, pp. 28–97.
Feyerabend, P. (1978). Science in a Free Society. London: New Left Books.
Flanagan, M., Howe, D., and Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. Van den Hoven and J. Weckert (Eds.), *Information Technology and Moral*

Philosophy (Cambridge Studies in Philosophy and Public Policy, pp. 322-353). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511498725.017

Floridi, L. and Sanders, J.W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14(3) 349-379.

Floridi, L. (2013). Distributed morality in an information society. Sci Eng Ethics (2013) 19: 727-743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. Fresco, N. and Primiero, G. (2015). On malfunctioning software. *Synthese.* 192(4): 1199–1220.

Floridi L. and Taddeo M. (2016). What is data ethics? Phil. Trans. R. Soc. A, Volume 374, Issue 2083. doi:10.1098/rsta.2016.0360. Available at SSRN: https://ssrn.com/abstract=2907744

Friedman, B. Nissenbaum, H. (1996). Bias in Computer Systems. ACM Transactions on Information Systems, 330-347.

Fukuyama, F. (1996). *Trust: The Social Virtues and The Creation of Prosperity*. New York, NY: The Free Press.

Gambetta, D. (1988). Can We Trust Trust? In *Trust. Making and Breaking Cooperative Relations*, edited by Diego Gambetta, 213-237. Oxford: Basil Blackwell.

Goldenberg, M. (2016). Public Misunderstanding of Science? Reframing the Problem of Vaccine Hesitancy. *Perspectives on Science.* 24(5):552-581.

Govier, T. (1997). *Social Trust and Human Communities,* Montreal and Kingston: McGill-Queen's University Press.

Govier, T. (1993). Self-Trust, Autonomy, and Self-Esteem. *Hypatia 8.1*, 99-120.

Gluckman, Sir Peter (2017). Science advice: A bastion against the post-truth/ post-trust torrent? Keynote address to the Annual Conference of the Joint Research Centre of the European Commission, Brussels, Available at: http://www.pmcsa.org.nz/wp-content/uploads/17-09-26-European-Commission-Joint-Research-Centre.pdf

Grandjean. P. Ozonoff, D. (2013). Transparency and translation of science in a modern world. *Environ Health* 12:70; doi:10.1186/1476-069X-12-70.

Hansen, C. Tosik, M. Goossen, G. C. Li, L. Bayeva, F. Berbain, and M. Rotaru (2015). How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence* / Symposium: Machine Learning Nijmegen.

Hardin, R. (1996). Trustworthiness, *Ethics*, 107: 26–42.

Hardin, R. (2002). *Trust and trustworthiness.* New York: Russell Sage Foundation.

Hardin, R. (2006). *Trust* Malden, MA: Polity Press.

Harding, S. (1991). *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Ithaca: Cornell University Press.

Hardwig, J. (1991). The role of trust in knowledge. *Journal of Philosophy*, 88(12), 693 - 708. Retrieved from http://search.proquest.com/docview/218173404

Hassan, S. and De Filippi, P. (2017). The Expansion of Algorithmic Governance: From Code is Law to Law is Code. Field Actions Science Reports: The Journal of Field Actions. Special issue 17: Artificial Intelligence and Robotics in the City. Open Edition Journals. Available at SSRN: https://ssrn.com/abstract=3117630

Hawley, K. (2017). Trustworthy Groups and Organizations. In *The Philosophy of Trust*. :Oxford University Press. (MLA)

Hildebrandt, M. (2008). Defining Profiling: A New Type of Knowledge?. In: Hildebrandt M., Gutwirth S. (eds) *Profiling the European Citizen.* Springer, Dordrecht 17–45. Available at: https://link.springer.com/chapter/10.1007/978-1-4020-6914-7_2

Holstein, K. Wortman Vaughan, J. H. Daumé, III, M. Dudik, and H. Wallach. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. New York: ACM, DOI: https://doi.org/10.1145/3290605.3300830

Holter, B. (2014). The Epistemic Significance of Values in Science. University of Calgary.

Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy* 72(1): 63–76.

Horel, S. and Bienkowski, B. (2013). Special report: scientists critical of EU chemical policy have industry ties. *Environmental Health News*. Available at: www.environmentalhealthnews.org/ehs/news/2013/eu-conflict)

Horsburgh, H. J. N. (1960). The Ethics of Trust. *Philosophical Quarterly.* 10(41): 343–354.

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

Inteman, K. (2017). Feminism, Values, and the Bias Paradox: Why Value Management Is Not Sufficient. 130-144. in C*urrent Controversies in Values in Science* eds. Elliot, K. and Steel, D. New York: Routledge

Intemann, K. and de Melo-Martín, I. (2014). Are there limits to scientists' obligations to seek and engage dissenters? *Synthese* 191(12): 2751-2765. doi:10.1007/s11229-014-0414-5
Ipsos MORI Veracity Index. Trust in Professions (2017). Available at: https://www.ipsos.com/ipsos-mori/en-uk/trust-professions-long-term-trends?view=wide
Jacynycz, V. Calvo, A. Hassan, S. Sánchez-Ruiz, A. A. (2016). Betfunding: A Distributed Bounty-Based Crowdfunding Platform over Ethereum. In: Omatu S. et al. (eds) Distributed Computing and Artificial Intelligence, 13th International Conference. Advances in Intelligent Systems and Computing, (474):403–411. Springer. ISSN: 2194-5357.
Jasanoff, S. (2016). *The Ethics of Invention - Technology and the Human Future.* New York: W. W. Norton and Company
John, S. (2015). Inductive risk and the contexts of communication. *Synthese*. 192(1): 79-96. doi:10.1007/s11229-014-0554-7
Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, 107: 4–25.
Jones, K. (2012). Trustworthiness. *Ethics*, 123(1): 61–85.
Jones, K. (2013). Distrusting the Trustworthy in *Reading Onora O'Neill*, ed. Archard, Deveaux, Manson, and Weinstock, London: Routledge: 186-199.
Jones, K. (2017). But I Was Counting On You! In *The Philosophy of Trust*.: Oxford University Press. Retrieved 7 Jun. 2017, from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198732549.001.0001/acprof-9780198732549-chapter-6.
Kahan, D. M. (2017). Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition. Cultural Cognition Project Working Paper Series No. 164; Yale Law School, Public Law Research Paper No. 605; Yale Law and Economics Research Paper No. 575. Available at SSRN: https://ssrn.com/abstract=2973067 or http://dx.doi.org/10.2139/ssrn.2973067
Kahan, D. M. (2013). Ideology, Motivated Reasoning, and Cognitive Reflection: An Experimental Study (November 29, 2012). Judgment and Decision Making, 8, 407-24 (2013); Cultural Cognition Lab Working Paper No. 107; Yale Law School, Public Law Research Paper No. 272. Available at SSRN: https://ssrn.com/abstract=2182588 or http://dx.doi.org/10.2139/ssrn.2182588
Kahneman, D. (2013). *Thinking, Fast And Slow.* London: Penguin Books
Katz, I. (2017). Have we fallen out of love with experts? *BBC News*. Available at:

https://www.bbc.com/news/uk-39102840

Kitcher, P. (2011). *Science in a Democratic Society*. New York: Prometheus Books.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication and Society* 20(1): 14–29.

Kourany, J. A. (2010). *Philosophy of science after feminism*. Oxford: Oxford University Press.

Kraemer, F. van Overveld, K. and Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.

Kuhn, T. S. (1962 [1970]). *The Structure of Scientific Revolutions*, Second edition, Chicago: University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, Value Judgment, and Theory Choice, in *The Essential Tension. Selected Studies in Scientific Tradition and Change*, Chicago: University of Chicago Press: 320–339.

Kurt, A. Semler, L. Jacoby, J.L. Johnson, M.B. Careyva, B.A. Stello, B. Friel, T. Knouse, M.C. Kincaid, H. and Smulian, J.C. (2016) Racial Differences Among Factors Associated with Participation in Clinical Research Trials. *Journal of Racial and Ethnic Health Disparities*: 1-10.

Lacey, H. (1999). *Is Science Value-Free? Values and Scientific Understanding.* London: Routledge.

Lacy, M. E. Wellenius, G. A. Sumner, A. E. Correa, A. Carnethon, M. R. Liem, R. I. Wilson, J. G. Saks, D. B. Jacobs D. R. Jr. Carson, A. P. Luo, X. Gjelsvik, A. Reiner, A. P. Naik, R. P. Liu, S. Musani, S. K.Eaton, C. B. Wu, W. C. (2017). Association of Sickle Cell Trait With Hemoglobin A1c in African Americans. *JAMA* 317, (5), 507-515.

Lambrecht, A. and Tucker, C. E. (2016). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260 .

Laudan, L. (2004). The Epistemic, the Cognitive, and the Social. In Machamer P. and Wolters G. (Eds.), *Science Values and Objectivity* Pittsburgh, Pa: University of Pittsburgh Press: 14-23. doi:10.2307/j.ctt5vkg7t.5

Levy, K. E. C. and Johns, D. M. (2016). When open data is a trojan horse: The weaponization of transparency in science and governance. *Big Data and Society. 3*(1) 2053951715621568. doi:10.1177/2053951715621568

List, C. and Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
Longino, H. E. (1990). *Science As Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, N.J.: Princeton University Press.
Longino, H. E. (1996). Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy, in *Feminism, Science and the Philosophy of Science*, L.H. Nelson and J. Nelson (eds.), Dordrecht: Kluwer, 39-58.
Longino, H. E. (2002). *The Fate of Knowledge*. Princeton, N.J.: Princeton University Press.
Lum, K.and Isaac, W. (2016). To predict and serve? *Significance* 13, No. 5 14-19.
Lycan, W. (1985). Epistemic Value. Synthese, 64(2), 137-164.
Mance, H. (2016). Britain has had enough of experts, says Gove. Financial Times. Available at: https://www.ft.com/content/3be49734-29cb-11e6-83e4-abc22d5d108c
Mazoue, J.G. (1990). Diagnosis without doctors. *Journal of Medicine and Philosophy* 15(6): 559–579.
Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City (May 16, 2018), Retrieved from: http://www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by
McGeer, V. (2008). Trust, Hope, and Empowerment. *Australasian Journal of Philosophy* 86(2): 237–254.
McLeod, C. (2002). *Self-Trust and Reproductive Autonomy*. Cambridge, MA: MIT Press.
McLeod, C. (2015). Trust, *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.), Available at: https://plato.stanford.edu/archives/fall2015/entries/trust/.
McMullin, E. (1982). Values in Science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*: 3–28.
McMullin, E. (2013). The Virtues of a Good Theory, in *The Routledge Companion to Philosophy of Science,* (ed.) Curd, M. and Psillos, Routledge Handbooks Online
Merton, Robert K. (1938). Science and the Social Order. *Philosophy of Science* 5, 3: 321–337.
Mitchell, S. (2004). The Prescribed and Proscribed Values in Science Policy. In Machamer P. and Wolters G. (Eds.), *Science Values and Objectivity* Pittsburgh. Pa: University of Pittsburgh Press: 245-255. doi:10.2307/j.ctt5vkg7t.16

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society, 3*(2), doi:10.1177/2053951716679679

Morrison, M. (2014). Values and Uncertainty in Simulation Models. *Erkenntnis* 79:939–959.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system, Available at: http://bitcoin.org/bitcoin.pdf

Nalisnick, E. Mitra, B. Craswell, N. and R. Caruana, R. (2016). Improving document ranking with dual word embeddings. *WWW'16 Companion.*
DOI: http://dx.doi.org/10.1145/2872518.2889361

Newell, S. and Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems* 24(1): 3–14.

Nickel, P. J. (2007). Trust and Obligation-Ascription. *Ethical Theory and Moral Practice* 10(3): 309–319.

Nissenbaum, H. (1998). Values in the Design of Computer Systems. Computers in Society, 38-39.

Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Enforce Racism* NYU Press.

Okruhlik, K. (1994). Gender and the Biological Sciences. *Canadian Journal of Philosophy*. 24:sup1, 21-42, DOI: 10.1080/00455091.1994.10717393

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge, New York: Cambridge University Press.

O'Neill, O. (2002b). *A Question of Trust: The BBC Reith Lectures.* Cambridge: Cambridge Univ Pr.

O'Neill, O. (2013). Responses in *Reading Onora O'Neill*, ed. Archard, Deveaux, Manson, and Weinstock, London: Routledge: 219–39.

O'Neill, O. (2014). Trust, Trustworthiness, and Accountability. In *Capital Failure: Rebuilding Trust in Financial Services*. Oxford: Oxford University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

Oreskes, N. and Conway, E. M. (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming.* New York: Bloomsbury.

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from You*. New York: The

Penguin Press.
Parker W. (2014).Values and Uncertainties in Climate Prediction, Revisited *Studies in History and Philosophy of Science.* 46:24–30.
Pasquale, F. (2016). *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, MA, US: Harvard University Press.
Pettit, P. (1995). The Cunning of Trust. *Philosophy and Public Affairs* 24: 202–225.
Pielke, R. (2007). *The Honest Broker: Making Sense of Science in Policy and Politics.* Cambridge University Press.
Pinto, M. F. (2014). Philosophy of science for globalized privatization. Uncovering some limitations of critical contextual empiricism. *Studies In History And Philosophy Of Science*. 4710-17. doi:10.1016/j.shpsa.2014.03.006
Pinto, M. F. (2015). Tensions in agnotology: Normativity in the studies of commercially driven ignorance. *Social Studies of Science.* 45(2): 294-315.
Popper, K. R. (1934 [2002]), *Logik der Forschung*, Berlin: Akademie Verlag. English translation as *The Logic of Scientific Discovery*, London: Routledge.
Popper, K. R. (1972*). Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.
Potter, N. N. (2002). *How Can I be Trusted? A Virtue Theory of Trustworthiness,* Lanham, Maryland: Rowman and Littlefield.
Proctor, R. N. (1995). *Cancer Wars: How Politics Shapes What We Know and Don't Know about Cancer.* New York: Basic Books.
Proctor, R. N. (1996). The anti-tobacco campaign of the Nazis: A little known aspect of public health in Germany:1933–45. *British Medical Journal* 313: 1450–1453.
Proctor, R. N. (1999). *The Nazi War on Cancer.* Princeton, NJ: Princeton University Press.
Proctor, R. N. (2008). Agnotology: A missing term to describe the cultural production of ignorance (and its study). In: Proctor, R. N. and Schiebinger, L. (eds) *Agnotology: The Making and Unmaking of Ignorance.*1–33. Stanford: Stanford University Press.
Proctor, R. N. (2011). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition.* Berkeley: University of California Press.
Proctor, R. N. and Schiebinger, L. (eds) (2008). *Agnotology: The Making and Unmaking of Ignorance.* Stanford: Stanford University Press.
Raymond, A. H. (2015). The Dilemma of Private Justice Systems: Big Data Sources, the Cloud

and Predictive Analytics, 35 *Northwestern Journal of International Law and Business*

Reisman, D. Schultz, J. Crawford, K. Whittaker, M. (2018). Algorithmic Impact Assessments: A practical Framework for Public Agency Accountability. Available at: https://ainowinstitute.org/aiareport2018.pdf

Reiss, J. and Sprenger, J. (2017). Scientific Objectivity. *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.) Available at: https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/

Rooney, P. (2017). The Borderlands between Epistemic and Non-Epistemic Values, in C*urrent Controversies in Values in Science* eds. Elliot, K. and Steel, D. New York: Routledge

Rubel, A. and Jones, K. M. L. (2016). Student privacy in learning analytics: An information ethics perspective, *The Information Society*, 32:2, 143-159, DOI: 10.1080/01972243.2016.1130502

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*. 20(1), 1-6.

Ruokonen, F. (2013). Trust, Trustworthiness, and Responsibility. *Trust: Analytic and Applied Perspectives*. Makaela, P. and Townley, C. (eds.) New York: Rodopi, 1-14.

Ruphy, S. (2006). "Empiricism all the way down": a defense of the value-neutrality of science in response to Helen Longino's contextual empiricism. *Perspectives on Science* 14(2), 189-214.

Sarewitz, D. (2016). Saving Science. *The New Atlantis*. (49):4-40.

Schwab, K. (2017). *The Fourth Industrial Revolution.* New York: Crown Business

Shackelford, S.J. and Raymond, A.H. (2014). Building the virtual courthouse: Ethical considerations for design, implementation, and regulation in the world of Odr. *Wisconsin Law Review* (3): 615–657.

Simon, J. (2010). The entanglement of trust and knowledge on the Web. *Ethics and Information Technology* 12(4): 343–355.

Sloman S. and Fernbach P. (2017). *The Knowledge Illusion, Why We Never Think Alone.* New York: Riverhead Books.

Smallman, M. (2018). Science to the rescue or contingent progress? Comparing 10 years of public, expert and policy discourses on new and emerging science and technology in the United Kingdom. *Public Understanding of Science* 27 (6) 655-673

Smith, T. W. and Son, J. (2013).Trends in Public Attitudes about Confidence in Institutions. General Social Survey 2012 Final Report (NORC at the University of Chicago, 2013).

Steel, D. (2016). Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk, *Perspectives on Science*, 24(6):696-721.

Steele, K. (2012). The Scientists Qua Policy Advisors Makes Value Judgements. *Philosophy of Science* 79 (5): 893-904.

Suberg, W. (2016). World Economic Forum: 'DLT' Blockchains Are the Future," Bitcoin News [Online]. Available: https://news.bitcoin.com/world-economic-forum-blockchain/

Swan, M. (2015). *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Media

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. Minds and Machines 20(2): 243–257.

Taddeo, M. and Floridi, L. (2015). The debate on the moral responsibilities of online service providers. Science and Engineering Ethics 1–29.

Taddeo, M. and Floridi, L. (2018). How AI can be a force for good. Science 361 (6404), 751-752, DOI: 10.1126/science.aat5991

Townley, C. and Garfield, J. L. (2013). Public Trust. Trust: Analytic and Applied Perspectives, P. Makela, P. and Townley, C. (eds.), Amsterdam: Rodopi Press

Turilli, M. and Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112

Tutt, A. (2016). An FDA for Algorithms. 69 Admin. L. Rev. 83 (2017). Available at SSRN: https://ssrn.com/abstract=2747994 or http://dx.doi.org/10.2139/ssrn.2747994

Uslaner, E. M. (1999). Democracy and Social Capital in *Democracy and Trust.* (ed.) Warren, M.E.: Cambridge University Press:121-150.

Vellido, A, Martın-Guerrero, J. D. and Lisboa, P. J. G. (2012). Making machine learning models interpretable. In: *ESANN 2012 proceedings*, Bruges, Belgium. 163–172.

Vigna P. and Casey, M. J. (2018). *The Truth Machine: The Blockchain and the Future of Everything*. New York: St. Martin's Press

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

Welch, S. (2013). Transparent trust and oppression. *Critical Review of International Social and Political Philosophy*. 16(1): 45–64.

Wilholt, T. (2009). Bias and Values in Scientific Research. *Studies in History and Philosophy of Science.* 40 (1): 92-101.

Wilholt, T. (2013). Epistemic Trust in Science, *The British Journal for the Philosophy of Science*. 64(2): 233-253.

Weinstock, D. (2013). Trust in Institutions. in *Reading Onora O'Neill*, ed. Archard, Deveaux, Manson, and Weinstock, London: Routledge: 199–218.

Wynne, B. (1996), May the Sheep Safely Graze? A Reflexive View of the Expert-Lay Knowledge Divide, in: Lash, S. et al. (eds.), *Risk, Environment, and Modernity. Towards a New Modernity, London*, 44-83

Wynne, B. (2006). Public engagement as a means of restoring public trust in science-hitting the notes, but missing the music? *Community Genetics.* 9(3):211-220.

Wylie, A. (2014). Community-Based Collaborative Archeology. *Philosophy of Social Science: A New Introduction.* (eds.) Cartwright, N. and Montuschi, E. Oxford University Press. 68-82.

Zarsky, T. Z. (2013). Transparent Predictions. *University of Illinois Law Review* (4): 1503-1569

Zarsky T. Z. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology and Human Values* 41(1): 118–132.

Curriculum Vitae

Marina Sonora graduated Philosophy and Croatian Language and Literature at the Faculty of Humanities and Social Sciences at the University of Zagreb in 2006. She was awarded Doctoral Fellowship at the University of Alberta, Canada, at the Wirth Institute for Austrian and Central European Studies during the 2016/2017 academic year. In 2013 she was awarded post graduate summer fellowship on Philosophy and it's relation to contemporary Science and Technology, by the Foundation of Hellenic World in Athens. Her further professional education included three months Massachusetts Institute of Technology Professional Education on the topic Digital Transformation: From AI and IoT to Cloud, Blockchain, and Cybersecurity in 2018/2019. During the 2014/2015 she attended several seminars on Responsible Research and Innovation, in Edinburg, on Open Science at FECYT, Madrid, Spain, on Interdisciplinary challenges.

Her recent papers and presentations include *Can Algorithms be Trustworthy*, *Synthesys Philosophica, 2019, Trustworthiness of science in the nexus between science, society and policy*, presented at Ryerson University, Canada, 2017, *Trustworthiness of science in the nexus between science, society and policy*, presented at the University of British Columbia, Vancouver, Canada, *Trustworthy science for policy: avoiding biases and misleading information* presented at the University of Alberta, Canada, 2017.

Marina works at ATG/Cognizant in the field of Information Technology and Services. Previously she worked as a Senior Expert Advisor in the Research and Innovation Program, Horizon 2020 at the Agency for Mobility and EU Programs. During that time she was participating in two Horizon 2020 projects, appointed as a Program Committee Member at the European Commission for the Societal Challenge 6 and Expert for the Science with and for Society, she lead over 20 workshops and info days, organized and participated in more than 20 workshops, events, and international conferences.