



# Kevin Baum

Martin-Luther-Straße 28, 66111 Saarbrücken

□ (+49) 173-7888-380 | ☎ uds@kevinbaum.de | 🏠 kevinbaum.de

| LinkedIn kevin-baum-55999580 | Google Scholar

DEKANAT DER PHILOSOPHISCHEN FAKULTÄT DER UNIVERSITÄT DES SAARLANDES  
REFERAT FÜR VERWALTUNG UND AKADEMISCHE VERFAHREN, FRAU HEIKE BRÜCKNER  
CAMPUS SAARBRÜCKEN, GEBÄUDE B3 1  
66123 SAARBRÜCKEN

March 28, 2024

## **Professor (W3) in Ethics of Digitalization (Reference no. W2428)**

Dear Professor Klakow, dear members of the selection committee,

You are looking for an experienced researcher at home in moral philosophy *and* computer science who explores the connections between the two. I am such a researcher and would like to apply.

My heart beats for the ethics of digitalization. I complemented computer science with philosophy, achieving degrees and remaining active in both disciplines. At *Saarland University*, I had the opportunity to develop an ethics course tailored to computer science students. My efforts culminated in the award-winning lecture series *Ethics for Nerds*, which has been going strong since 2016. I was also a driving force behind and in the large and successful interdisciplinary research project *Explainable Intelligent Systems* (EIS, 2019–2024). By moving to the *German Research Center for Artificial Intelligence* (DFKI) at the beginning of 2023, I seized the chance to research cutting-edge AI and take on new responsibilities. Yet, even as deputy head of the DFKI research department *Neuro-Mechanistic Modeling* and head of the new *Center for European Research in Trusted AI* (CERTAIN), the ethical dimensions of AI utilization and the pursuit of trustworthy AI remain central to my work.

While my dissertation was on foundational issues in consequentialist ethics, I have amassed an entirely independent body of publications on the ethics of digitalization. This corpus coincides with a wide range of activities listed in my CV: extensive teaching, numerous talks in academic settings, roles encompassing wide-ranging briefs (such as finances, staff management, and strategic planning), acquisition of funding, organization of conferences, legislative advising, and public communication of science. I built networks and spearheaded interdisciplinary projects; I communicated effectively with regulators, industry professionals, and businesses; and I proved my ability to assemble highly motivated teams that excel and thrive on collaboration.

Both in my research and teaching I have investigated how the explainability of algorithmic decision-making relates to normative and social desiderata. I am deeply concerned with the societal risks posed by opacity, the uncovering of unfair biases in models, and the philosophical questions surrounding AI alignment. Reflecting on my contributions and future ambitions, I see numerous opportunities for collaboration with the *Philosophy Department* and *Faculty P* at large, including the envisaged *Center for Digital Humanities*.

Opportunities to join forces with computer science abound. My research on model-agnostic local explainability methods, designed to bolster human oversight, aligns seamlessly with the *Transregional Collaborative Research Centre 248: Foundations of Perspicuous Software Systems*. My work related to theoretical ethics and game theory could lead the path to a philosophically well-informed integration of methods from safe ML into applied machine ethics – for instance, in the context of reinforcement learning. Furthermore, I have a thorough understanding of, and network within, the research landscape at the *Saarland Informatics Campus* and CISPA. It would be fruitful to cooperate, for example, with the *Interdisciplinary Institute for Societal Computing* of Professors Ingmar Weber and Daniela Braun, with the group of Prof. Isabel Valera, and within the DFKI across different locations. For detailed proposals, please see the attached research.

Building bridges between philosophy and computer science has been my passion for over a decade. I would be delighted to present my ideas and plans to the selection committee.

With best regards

MARCH 28, 2024

KEVIN BAUM · COVER LETTER

Attached:

- academic curriculum vitae,
- copies of my degree certificates,
- a full list of publications,
- full list of third-party funding,
- research plan,
- a teaching statement,
- full-text copies of my most important publications,
- a list of (more than) three academic references, and
- a statement on habilitation equivalence.

# Kevin Baum, M.Sc., M.A.

 Martin-Luther-Str. 28, 66111 Saarbrücken, Saarland, Germany

 28.07.1986, Schwetzingen

 +49 173 7888 380

 academia@kevinbaum.de

 <https://kevinbaum.de>

 Scholar

 LinkedIn

 → That symbol indicates items in this CV and elsewhere in this application which I propose to be cumulatively equivalent to a habilitation; see also the separate document »Habilitation Equivalence«.

## (Academic) Employment History (since the end of my Master's studies)

- |   |   |
|---|---|
|  12/2023 – today |  <b>Head of the Center for European Research in Trusted AI (CERTAIN).</b><br>German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Saarland.   |
|  01/2023 – today |  <b>Deputy Head and Manager of the Research Department for Neuro-Mechanistic Modeling (NMM).</b><br>German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Saarland.  |
| 02/2015 – 12/2022   |  <b>Research Assistant and Lecturer.</b><br>Group for Dependable Systems and Software (Prof. Hermanns), since 2018 in project <i>Explainable Intelligent Systems</i> , funded by VolkswagenStiftung, Saarland University (UdS), Saarbrücken, Saarland.<br> <b>Research Assistant and Lecturer.</b><br>Chair for Practical Philosophy (Prof. Wessels, Prof. Fehige), since 2018 in project <i>Explainable Intelligent Systems</i> , funded by VolkswagenStiftung, Saarland University (UdS), Saarbrücken, Saarland. |

## Education

- |      |  |                        |
|------|--|------------------------|
| 2024 |  <b>Doctor of Philosophy</b> at TU Dortmund                                 | <i>summa cum laude</i> |
|      | Thesis title: <i>Doing Wrong with Others – Multi-Agent Consequentialism as a Solution for the Collective Action Problem</i> . Defended, publication pending.   |                        |
| 2014 |  <b>M.A. Philosophy</b> at Saarland University                              | 1.1                    |
|      | Thesis title: <i>Vom Bezug singulärer Terme in Aussagen über propositionale Einstellungen</i> .  |                        |
| 2013 |  <b>M.Sc. Computer Science</b> at Saarland University                       | 1.4                    |
|      | Thesis title: <i>GPGPU-gestützte diffusionsbasierte naive Videokompression</i> .   |                        |
| 2011 |  <b>B.Sc. Computer Science, (Minor: Mathematics)</b> at Saarland University | 2.0                    |
|      | Thesis title: <i>Stützstellenauswahl für diffusionsbasierte Bildkompression unter Berücksichtigung einer Quadrikel-Substruktur-Restriktion</i> .               |                        |
| 2006 |  <b>Abitur</b> at Gymnasium Süderelbe, Hamburg                              | 1.5                    |

## Awards and Achievements

- |  |   |
|--|---|
|  2020 |  <b>Hochschulperle of the Year</b> for lecture »Ethics for Nerds«, German Stifterverband.            |
|  2019 |  <b>Hochschulperle of the Month (January)</b> for lecture »Ethics for Nerds«, German Stifterverband. |
| 2011, 2012   |  <b>Deutschlandstipendium</b>  |

## Teaching

- |                  |   |
|------------------|---|
| Summer 2024      | ■ » <b>Practical AI Ethics</b> « (upcoming) with Sarah Sterz. KI Campus.<br><i>(online course, in preparation)</i>  |
| Summer 2023      | ■ » <b>AI for the Social Good</b> « ( <b>Seminar</b> ) with Dr. Gerrit Großmann, Lisa Dargasz, Sarah Sterz, Prof. Verena Wolf. Saarland Informatics Campus, UdS.<br>7 ECTS      |
|                  | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Sarah Sterz, Prof. Holger Hermanns. Saarland Informatics Campus, UdS.<br>6 ECTS   |
| Summer 2022      | ■ » <b>Computer Ethics</b> « ( <b>Lecture</b> ) with Sarah Sterz. LL.M. <i>Informationstechnologie und Recht</i> , UdS.<br>3 ECTS   |
|                  | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Sarah Sterz, Prof. Holger Hermanns. UdS.<br>6 ECTS  |
| Winter 2021/2022 | ■ » <b>Ethische Fragen des Technologieeinsatzes</b> « ( <b>Seminar</b> ). Bucerius Law School, Hamburg.<br>2 ECTS   |
| Summer 2021      | ■ » <b>Computer Ethics for IT &amp; Law</b> « ( <b>Lecture</b> ) with Sarah Sterz. LL.M. <i>Informationstechnologie und Recht</i> , UdS.<br>3 ECTS                              |
|                  | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Sarah Sterz, Prof. Holger Hermanns. Saarland Informatics Campus, UdS.<br>6 ECTS   |
| Winter 2020/2021 | ■ » <b>Ethische Fragen des Technologieeinsatzes</b> « ( <b>Seminar</b> ). Bucerius Law School, Hamburg.<br>2 ECTS   |
| Summer 2020      | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Sarah Sterz, Prof. Holger Hermanns, Saarland Informatics Campus, UdS.<br>6 ECTS   |
| Summer 2019      | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Sarah Sterz, Prof. Holger Hermanns, Saarland Informatics Campus, UdS.<br>6 ECTS   |
| Winter 2018/2019 | ■ » <b>Computer sagt: „wahrscheinlich“. Wie KI und Algorithmen unsere Welt verändern</b> « ( <b>Seminar</b> ). College of Fine Arts (Universität der Künste), Berlin.<br>2 ECTS |
|                  | ■ » <b>Weapons of Math Destruction – Wie Big Data Ungerechtigkeit fördert und Demokratie gefährdet</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS              |
| Summer 2018      | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Prof. Holger Hermanns. Saarland Informatics Campus, UdS.<br>6 ECTS  |
|                  | ■ » <b>Derek Parfits Praktische Philosophie in Reasons &amp; Persons</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS  |
| Winter 2017/2018 | ■ » <b>Das Problem gemeinschaftlichen Handels</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS   |
| Summer 2017      | ■ » <b>Moral und Algorithmen</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS  |
|                  | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Prof. Holger Hermanns. Saarland Informatics Campus, UdS.<br>6 ECTS  |
| Winter 2017/2018 | ■ » <b>Von der Entscheidungstheorie zum Konsequentialismus</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS  |
| Summer 2016      | ■ » <b>Extending Morals: Moralisch handelnde Roboter, Roboter moralisch behandeln</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.<br>6 ECTS                               |

## Teaching (continued)

- |                  |   |               |
|------------------|---|---------------|
| Winter 2015/2016 | ■ » <b>Ethics for Nerds</b> « ( <b>Lecture</b> ) with Prof. Holger Hermanns. Saarland Informatics Campus, UdS.  | <i>6 ECTS</i> |
| Summer 2015      | ■ » <b>Was kommt jetzt? Die Zukunft der Menschheit, Transhumanismus und die technologische Singularität</b> « ( <b>Seminar</b> ). Philosophy Department, UdS. | <i>6 ECTS</i> |
| Summer 2013      | ■ » <b>Es macht doch keinen Unterschied! Oder: Was wäre, wenn jeder das täte?</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.                           | <i>6 ECTS</i> |
|                  | ■ » <b>Ethik für Nerds</b> « ( <b>Proseminar</b> ) with Prof. Holger Hermanns. Saarland Informatics Campus, UdS.  | <i>5 ECTS</i> |
|                  | ■ » <b>Prädikatenlogik erster Stufe</b> « ( <b>Seminar</b> ). Philosophy Department, UdS.   | <i>6 ECTS</i> |

## Supervision and Review of Student Theses, Student Advising

### Independently Supervised and Reviewed Bachelor's and Master's Theses

- |      |  |
|------|--|
| 2024 | ■ [A thesis on machine ethics, upcoming]. Master's Thesis of Lisa Dargasz ( <b>M.Sc.</b> ).  |
| 2023 | ■ <i>Outputs as Evidence: Exploring the Explanatory Role of XAI Methods through Thomas Bartelborth's Theory on Nomic Patterns</i> . Bachelor's Thesis of Lena Marie Budde ( <b>B.A.</b> ). |
| 2018 | ■ <i>Another Take on Newcomb's Problem</i> . Bachelor's Thesis of Tim Dirk Metzler ( <b>B.A.</b> ).  |

### Co-Supervised Bachelor's and Master's Theses, Advisory Service

- |           |   |
|-----------|---|
| 2018-2023 | ■ <i>Building Bridges for Better Machines: From Machine Ethics to Machine Explainability and Back</i> . Doctoral Thesis of Timo Speith ( <b>Dr. Phil.</b> ).    |
| 2020      | ■ <i>Gibt es Umstände, unter denen Abtreibung moralisch erlaubt ist?</i> State examination thesis of Alisha Monika Mateas ( <b>StEx</b> ).                      |
| 2018      | ■ <i>A Framework of Verifiable Machine Ethics and Machine Explainability</i> . Master's Thesis of Timo Speith ( <b>M.Sc.</b> ).                                 |
| 2017-2018 | ■ <i>Schuld der Massen – Kollektive Verantwortung und das Überdeterminiertheitsproblem</i> . Master's Thesis of Yannik Bast ( <b>M.A.</b> ).                    |
| 2017      | ■ <i>Klimawandel und Spieltheorie: Ist kooperatives Verhalten der Staaten in der Klimapolitik rational?</i> Bachelor's Thesis of Tessy Aulner ( <b>B.A.</b> ).  |
|           | ■ <i>Warum ich einen Unterschied mache. Zur Ethik individuellen Handelns im gemeinschaftlichen Handeln</i> . Bachelor's Thesis of Sarah Maurer ( <b>M.A.</b> ). |

### Co-Supervised Research Immersion Labs

- |      |   |
|------|---|
| 2024 | ■ Lena Marie Budde: [A research immersion lab on explainable AI and unfairness detection, upcoming].      |
| 2023 | ■ Janine Lohse: <i>Predicting and Manipulating Game Rules through Weight Analysis in Neural Networks?</i> |

## **Research Publications**

A full list of publications (and further academic activities) is provided as a separate part of this application.

## **Other Academic Activities (Selection)**

### **Renowned Workshops and Seminars (by Invitation Only)**

- 1** *Workshop on Artificial Intelligence at the UK-German Science, Innovation and Technology Dialogue*, Imperial College White City Campus, London, UK, March 12, 2024.
- 2** *Dagstuhl Seminar 23371: Roadmap for Responsible Robotics*, Leibniz Center for Informatics Schloss Dagstuhl, Sep 10 – Sep 15, 2023.  
(The actual *Roadmap for Responsible Roadmaps* as a joint Dagstuhl publication under the direction of Prof. Michael Fisher is currently being finalized.)
- 3** *Dagstuhl Seminar 23151: Normative Reasoning for AI*, Leibniz Center for Informatics Schloss Dagstuhl, Apr 10 – Apr 14, 2023.
- 4** *Dagstuhl Seminar 16222: Engineering Moral Agents – from Human Morality to Artificial Morality*, Leibniz Center for Informatics Schloss Dagstuhl, May 29 – Jun 03, 2016.

### **Academic Talks (Selection)**

- 1** K. Baum, *Beyond Technicalities: Positive Normative Choices and the Ethical Dimensions of Responsible AI System Design*, at the *1st Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISoLA 2023), Crete, Greece, Oct. 2023.  URL: <https://aisola.org/papers/baum.pdf>.
- 2** K. Baum, S. Biewer, H. Hermanns, et al., *Effective Human Oversight: Conditions and Implications of the Proposed EU AI Act from an Interdisciplinary Perspective*, at the *1st Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISoLA 2023), Crete, Greece, Oct. 2023.  URL: <https://aisola.org/papers/hermanns-lauber-roensberg-langer-baum-hetmank-sterz-meinel-biewer.pdf>.
- 3** N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer, *How Do We Assess System Trustworthiness? Introducing the Trustworthiness Assessment Model*, at the *1st Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISoLA 2023), Crete, Greece, Oct. 2023.  URL: <https://aisola.org/papers/schlicker-baum-uhde-sterz-hirsch-langer.pdf>.
- 4** K. Baum, *Machine Ethics ⇔ Machine Explainability*, guest lecture and expert panel participation on invitation by Rune Nyrup in context of the seminar »Designing artificial moral agents – could we, and should we?« within the program »AI Ethics and Society« at the *Centre for the Future of Intelligence* (CFLI), Cambridge, UK, Jan. 2022.
- 5** K. Baum, M. Langer, and S. Sterz, *The Role of XAI for Meta-Trust and Trust Propagation - Psychological and Philosophical Perspectives*, at the *2nd Workshop on Issues in XAI: Understanding and Explaining in Healthcare* at the *Centre for the Future of Intelligence* (CFLI), Cambridge, UK (postponed and moved online due to COVID-19), May 2021.  URL: <http://lcfi.ac.uk/news-and-events/events/explainable-ai-2-understanding-and-explaini/>.

- 6 K. Baum and S. Sterz, *Ethics for Nerds: Best Practices for Teaching Ethics to Computer Scientists*, at the *10th Annual Symposium Ethics in the Age of Smart Systems* (online), Apr. 2021. Ⓢ URL: <https://www.luc.edu/digitaletics/events/symposiaarchive/2021/>.
- 7 K. Baum, *New Ground for Consequentialism: How to Solve the Coordination Problem*, paper presentation at Thomas Schmidt's *Colloquium in Practical Philosophy* at the Humboldt University (HU), Berlin, Germany (online due to COVID-19), Jan. 2021.
- 8 K. Baum, *Verantwortung, Vertrauen, Rechte: Über die ethische Dimension erklärbarer KI*, at the workshop *Verantwortlichkeit digitalisierter Unternehmen – die ethischen und rechtlichen Auswirkungen des Einsatzes künstlicher Intelligenz*, Universität Salzburg, Salzburg, Austria, Nov. 2019. Ⓢ URL: <https://www.plus.ac.at/wp-content/uploads/2021/02/Programmentwurf.pdf>.
- 9 K. Baum, *Development or Regulation First*, at the *1st Workshop on Issues in XAI: Blackboxes, Recommendations, and Levels of Explanations*, Saarbrücken, Germany, Oct. 2019. Ⓢ URL: <https://explainable-intelligent.systems/workshop/>.
- 10 K. Baum, L. Kästner, and E. Schmidt, *Understanding Explainable AI: The EIS Project*, at the scientific colloquium of the »Science, Value, and the Future of Intelligence« project at the *Leverhulme Centre for the Future of Intelligence*, Cambridge, UK, Jul. 2019.
- 11 K. Baum and F. Bräuer, *Trusting Artificial Experts*, at the workshop *Evidence in Law and Ethics*, Jagiellonian University, Kraków, Poland, Apr. 2019. Ⓢ URL: <https://incet.uj.edu.pl/evidence-in-law-and-ethics>.
- 12 K. Baum and E. Schmidt, *Moral Harmony vs. Supervenience: A New Dilemma for Consequentialism*, at the *10th International Congress of the Society for Analytical Philosophy* (GAP 10), Cologne, Germany, Sep. 2018. Ⓢ URL: [https://gap10.de/wp-content/uploads/2018/09/Programmheft.GAP\\_.10.Final\\_.NEU\\_-.pdf](https://gap10.de/wp-content/uploads/2018/09/Programmheft.GAP_.10.Final_.NEU_-.pdf).
- 13 K. Baum, H. Hermanns, and T. Speith, *Towards a Framework Combining Machine Ethics and Machine Explainability*, 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology (CREST2018) as part of the 28th European Joint Conferences on Theory and Practice of Software (ETAPS 2018), Thessaloniki, Greece, Apr. 2018. Ⓢ URL: <https://www.react.uni-saarland.de/crest2018/>.
- 14 K. Baum, *Obligation Overboard? Decision-Theoretic Objective Consequentialism to the Rescue*, talk as part of a research stay at the *Munich Center for Mathematical Philosophy* (MCMP) at the invitation of Prof. Dr. Stephan Hartmann, Munich, Germany, Jul. 2017.
- 15 K. Baum, *What the Hack Is Wrong with Software Doping?* at the *7th International Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (ISOLA 2016), Corfu, Greece, Oct. 2016.
- 16 K. Baum and E. Schmidt, *Kagans 'Lösung' des Problems gemeinschaftlichen Handelns: Wie eine verborgene Annahme Kagans Ansatz entwertet*, poster presentation at the *9th International Congress of the Society for Analytical Philosophy* (GAP 9), Osnabrück, Germany, Sep. 2015. Ⓢ URL: <https://gap9.de/wp-content/uploads/2015/08/Programmheft-GAP.9-final.pdf>.
- 17 K. Baum, *Asking the Right Questions*, at the seminar *Engineering Moral Agents – from Human Morality to Artificial Morality* on Schloss Dagstuhl, Leibniz-Zentrum für Informatik, Wadern, Germany, May 2015. Ⓢ URL: <https://drops.dagstuhl.de/entities/document/10.4230/DagRep.6.5.114>.

## Upcoming, Invitation Accepted

- 1 K. Baum, *Erklärbarkeit und Effective Human Oversight im Art. 14 des EU AI Acts*, talk at the seminar *KI und der aufgeklärte Nutzer* of the *Law and AI Research Group (LARG)* at *Bucerius Law School*, Hamburg, Germany, Apr. 2024.
- 2 K. Baum, *On the Quest for Effectiveness in Human Oversight*, talk at the *Institute for Ethics in Technology* at *Hamburg University of Technology (TUHH)*, Hamburg, Germany, Apr. 2024.

## Presentations Aimed at Public Audience

### Talks and Panel Appearances (Selection)

- 1 K. Baum, *Ethik und KI: Eine Balance zwischen Fortschritt und Verantwortung?* talk and expert panel participation at the *Afterwork-Event „Business-KI-Balance“* @CFK/GymLodge event in context of the *European Digital Innovation Hub Saarland (EDIH)*, Spiesen-Elversberg, Germany, Mar. 2024.
- 2 K. Baum, *AI Safety and Security in Practice*, talk as part of the visit of the »Trade Mission to Saarland« of the *Luxembourg Chamber of Commerce* at the invitation of the *Saarland Ministry of Economics, Innovation, Digital Affairs and Energy*, Mar. 2024. ⚡ URL: <https://app.swapcard.com/event/mission-germany/exhibitors/RXZlbnRWaWV3XzY2Nzg3MQ==>.
- 3 K. Baum, *Künstliche Intelligenz im Betrieb: Ethische Fallstricke*, talk at »Technologiekonferenz« of the *Beratungsstelle für sozialverträgliche Technologiegestaltung e.V. (BEST)* and the *Arbeitskammer des Saarlandes*, Dec. 2023. ⚡ URL: <https://www.rechtsschutzsaal.de/aktuelles-termine/meldung-1/technologiekonferenz-best-ev-im-rechtsschutzsaal-bildstock>.
- 4 K. Baum, *KI – Fluch oder Segen?* Expert panel participation in the context of the »Der Fabulant« project of *modus|zad*, Dec. 2023. ⚡ URL: <https://modus-zad.de/publikation/blog/online-veranstaltung-ki-fluch-oder-segen/>.
- 5 K. Baum, *KI und deren Bedeutung für Wissenschaft und Hochschulen*, expert panel participation in the context of the »Studium Generale« at *Bucerius Law School*, Jun. 2023. ⚡ URL: <https://www.law-school.de/news-artikel/expertengespraech-zu-ki-und-die-bedeutung-fuer-wissenschaft-und-hochschulen>.
- 6 K. Baum, *Künstliche Intelligenz und Verantwortung*, keynote at the *Paul Fritzsche Stiftung Wissenschaftliches Forum*, Homburg, Germany, Dec. 2022. ⚡ URL: [https://www.uniklinikum-saarland.de/de/lehre/dekanat/wissenschaftliche\\_foren\\_gastvortraege](https://www.uniklinikum-saarland.de/de/lehre/dekanat/wissenschaftliche_foren_gastvortraege).
- 7 K. Baum, *AI. Friend, Foe or Fad?* Expert panel participation in the context of the *Booster Conference 2024* in Bergen, Norway, Apr. 2022. ⚡ URL: <https://2022.boosterconf.no/talk/panel-ai-friend-foe-or-fad/>.
- 8 K. Baum, *Algorithmen, Fake News & Fragmentierung*, talk at »Safer Internet Day 2021«, Feb. 2021. ⚡ URL: [https://www.onlinerlandsaar.de/wp-content/uploads/2021/01/safer\\_internet\\_day\\_2021\\_X4.pdf](https://www.onlinerlandsaar.de/wp-content/uploads/2021/01/safer_internet_day_2021_X4.pdf).
- 9 K. Baum, *Nudging, Manipulation, Profiling*, talk at »Aktionstage Netzpolitik & Demokratie 2020« of the *Zentralen für politische Bildung (ZpB)*, Nov. 2020. ⚡ URL: <https://www.youtube.com/watch?v=PM8r70hJZnI>.

- 10** K. Baum, *Moralische Hürden von Profiling und die Herausforderung der Erklärbarkeit*, talk at the event »Profiling 2.0« organized by the Thuringian State Commissioner for Data Protection and Freedom of Information (TLfDI), Oct. 2020.  URL: [https://www.onlinerlandsaar.de/wp-content/uploads/2021/01/safer\\_internet\\_day\\_2021\\_X4.pdf](https://www.onlinerlandsaar.de/wp-content/uploads/2021/01/safer_internet_day_2021_X4.pdf).
- 11** K. Baum, *Fünf ethische Herausforderungen im Zeitalter der digitalisierten Medizin*, keynote (»Festvortrag«) at the *Eröffnung des Fortbildungsjahres 2020 / 2021* of the *Ärztekammer des Saarlandes*, Homburg, Germany, Sep. 2020.  URL: <https://www.aerztekammer-saarland.de/index/news/News-20200824-Eroeffnung-Fortbildungsjah/>.
- 12** K. Baum, *Wenn Computer die Lebenszeit vorhersagen: Autonomie, Verantwortung und Erklärbarkeit*, talk at the »Hospizgespräch« at St. Jakobus Hospiz., Jun. 2019.  URL: [https://www.kinderhospizdienst-saar.de/uploads/media/18-05-28\\_Einladung\\_JUNI\\_01.pdf](https://www.kinderhospizdienst-saar.de/uploads/media/18-05-28_Einladung_JUNI_01.pdf).
- 13** K. Baum, *Aber warum, Computer? Erklär' es mir! Von rassistischen, undurchschaubaren künstlichen Intelligenzen, die vielleicht morgen schon über Ihr Leben bestimmen und von der Hilflosigkeit von Menschen in Schleifen*, talk at the »Futurologischer Kongress« of the *Stadttheater Ingolstadt.*, Jun. 2018.  URL: [https://theater.ingolstadt.de/fileadmin/doc/dokumentation\\_futurologischer\\_kongress/futurologischer\\_kongress\\_2018.pdf](https://theater.ingolstadt.de/fileadmin/doc/dokumentation_futurologischer_kongress/futurologischer_kongress_2018.pdf).

## (Co-)Organizer of Events (Selection)

- 1** *Automatisierte Aufmerksamkeit: Wie KI Journalismus, Öffentlichkeit und Meinungsbildung prägt*, second joint event of the *Landeszentrale für politische Bildung des Saarlandes* (LpB), the non-profit association *Algoright e.V.*, the *German Research Center for Artificial Intelligence* (DFKI) and the *Saarland State Media Authority* (LMS)., Nov. 2023.  URL: [https://eveeno.com/automatisierte\\_aufmerksamkeit](https://eveeno.com/automatisierte_aufmerksamkeit).
- 2** *Entmystifizierung eines Hypes: ChatGPT zum Anfassen*, first joint event of the *Landeszentrale für politische Bildung des Saarlandes* (LpB), the non-profit association *Algoright e.V.*, the *German Research Center for Artificial Intelligence* (DFKI) and the *General Studies Committee* (AStA) of *Saarland University.*, May 2023.  URL: [https://eveeno.com/ChatGPT\\_zum\\_anfassen](https://eveeno.com/ChatGPT_zum_anfassen).
- 3** *Synergien oder Konflikt? Forschung, Regulierung und Innovation im Spannungsverhältnis am Beispiel Künstlicher Intelligenz*, expert panel, May 2023.  URL: <https://www.saarnews.com/podiumsdiskussion-in-saarbruecken-kuenstliche-intelligenz-an-der-schnittstelle-von-wissenschaft-politik-und-wirtschaft/>.

## Upcoming, Invitation Accepted

- 1** K. Baum, *Vertrauenswürdige KI: Mehr als ein Buzzword*, research keynote at the *Journalismuspreis Informatik 2024* of the *Ministerium für Wirtschaft, Innovation, Digitales und Energie des Saarlandes* at the *Saarland Informatics Campus*, Saarbrücken, Germany, May 2024.
- 2** K. Baum, *Trusted AI – An Opportunity for German-French Startups*, expert panel participation at the *Viva Technology 2024*, Paris, France, May 2024.
- 3** K. Baum, *Trusted AI as a Driving Force for Societally Beneficial Innovations*, keynote at the *1st Industry Collaboration and Transfer Exchange – From Research to Market*, Saarbrücken, Germany, May 2024.

4

*KI und Bildung: Der Stand der Dinge im Frühjahr 2024*, third joint event of the *Landeszentrale für politische Bildung des Saarlandes* (LpB), the non-profit association *Algoright e.V.*, the *German Research Center for Artificial Intelligence* (DFKI) with, among others, the *Saarland Minister of Education* Christine Streichert-Clivot, Jun. 2024.



## Scientific Services & Science Infrastructure

### (Co-)Organizing of Workshops, & Special Tracks, Tutorials

- 2024
- **Special Track Co-Organizer** of »Responsible and Trusted AI: An Interdisciplinary Perspective« at the *2nd Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISOLA 2024), 30 October – 3 November, 2024, Crete, Greece.
  - **Tutorial Co-Organizer** (under review) of »Machine Ethics« at the *33rd International Joint Conference on Artificial Intelligence* (IJCAI 2024), 3 August – 9 August, 2024, Jeju, South Korea.
- 2023
- **Special Track Co-Organizer** of »Responsible and Trustworthy AI: Normative Perspectives on and Societal Implications of AI Systems« at the *1st Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISOLA 2024), 23 – 28 October 2023, Crete, Greece.
  - **Special Track Co-Organizer** of »Interdisciplinary Perspectives on XAI« at the *1st World Conference on eXplainable Artificial Intelligence* (xAI 2023), 26 – 28 July 2023, Lisboa, Portugal

### Program Committee Memberships

- 2024
- **Program Committee Member** of the *2nd Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISOLA 2024), 30 October – 3 November, 2024, Crete, Greece.
  - **Program Committee Member** of the special track on »Agents in Ethics« at the *21st European Conference on Multi-Agent Systems* (EUMAS 2024), 26 – 28 August 2024, Dublin, Ireland.
  - **Program Committee Member** of the *23rd International Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS 2024), 6 – 10 May 2024, Auckland, New Zealand.
- 2023
- **Program Committee Member** of the *1st Artificial Intelligence Symposium On Leveraging Applications of Formal Methods, Verification and Validation* (AISOLA 2023), 23 – 28 October 2023, Crete, Greece.

### Reviewer Activities for Academic Journals

- Philosophical Studies, Philosophy & Technology, Big Data & Society, Communications of the ACM

### Ethics Committees, Commissions, Advisory Functions

- 2024
- Member of the **delegation of Federal Minister of Education and Research Bettina Stark-Watzinger** for the high-level **UK-German Science, Innovation and Technology Dialogue**, Imperial College White City Campus in London, March 12.



## Scientific Services & Science Infrastructure (continued)

- 2023-\* **Ethics Advisor.** Ethics Team, German Research Center for Artificial Intelligence (DFKI).  
 DFKI Representative in the ***East Side Fab e.V. board*** (on behalf of Prof. Verena Wolf).
- 2021-\* Member of the **Ethical Review Board**. Faculty of Mathematics and Computer Science, Saarland University.
- 2020-2022 **Permanent expert on digital ethics** at the **Enquête Commission »Digitalisierung im Saarland – Bestandsaufnahme, Chancen und Maßnahmen«** of the Saarland State Parliament.
- 2019-2022 Member and Deputy Chairman of the »**Kommission für die Ethik sicherheitsrelevanter Forschung**«. Saarland University.
- 2018-2022 **Ethics Advisor.** Undisclosed Horizon 2020 Project (together with Sarah Sterz and Timo Speith).

## Miscellaneous Experience

### Further Experiences

- Media I have participated in various media productions, from regional outlets (SR (radio, including features), SZ) to national ones (DLF (Kultur), ARD) to online formats (podcasts on various topics, Youtube, heise online, Spektrum der Wissenschaft and other formats). Recently I was interviewed by Handelsblatt. I come from a journalism household and practically grew up with recording technology. I also became the UdS's contact person for *machine ethics, XAI, technology assessment, and the ethics of digitalization* as part of the *expert service for journalists*.
- Politics and Regulation I was involved in the preparation of statements for the state parliament on topics such as computer-aided procedures at the Saarland police in the area of accident investigation and the use of technology in the school context (*Bildungsausschuss*, March 2023), but also for the *Bundestag* (on the draft of an act on autonomous driving). In addition, my advice was taken up by the Saarland State Commissioner for Data Protection and Freedom of Information, and I was involved in background discussions with politicians in the Bundestag and the European Parliament on issues relating to the design and classification of two drafts of the AI Act.
- Startups, SMEs, Industry In the past, I have designed and organized various large-scale, paid consultations and workshops for clients like Villeroy & Boch and Accso. I also have start-up experience: I was part of an *EXIST-Gründungsstipendium* funded team (2016 to 2017). In the last years, I have built up reliable contacts with various medium-sized companies and industrial partners, including ZF and eurodata.

## Miscellaneous Experience (continued)

### Memberships

- since 2019       **Algoright e.V.** (non-profit/»gemeinnützig«)  
*founding member*
- since 2018       **Gesellschaft für Informatik (GI)**
-  **Universitätsgesellschaft des Saarlandes e.V.**
- since 2016       **Gesellschaft für Utilitarismusstudien (GUS)**  
*founding member*
- since 2015       **Gesellschaft für Analytische Philosophie (GAP)**

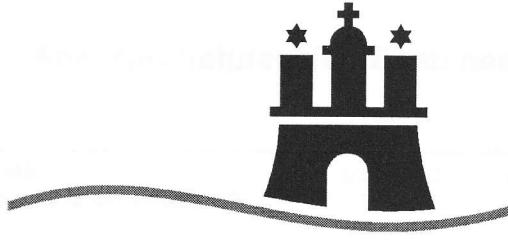
## Language Skills & Hobbies

### Languages

-  German (native), English (fluent)

### Hobbies

-  streetball, bouldering, hiking, reading (fiction), chess, cats



# FREIE UND HANSESTADT HAMBURG

## GYMNASIUM SÜDERELBE

### Z E U G N I S

### DER ALLGEMEINEN HOCHSCHULREIFE

**Kevin Baum**

geboren am 28.07.1986 in Schwetzingen

hat nach dem Besuch der gymnasialen Oberstufe die Abiturprüfung abgelegt.

Dem Zeugnis liegen zugrunde:

Die „Ausbildungs- und Prüfungsordnung zum Erwerb der allgemeinen Hochschulreife - APO-AH vom 22.07.2003“ (Hamburgisches Gesetz- und Verordnungsblatt Seite 275) in der jeweils geltenden Fassung,

die „Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II“ (Beschluss der Kultusministerkonferenz vom 07.07.1972) in der jeweils geltenden Fassung,

die „Vereinbarung über die Abiturprüfung der gymnasialen Oberstufe in der Sekundarstufe II“ (gemäß Vereinbarung der Kultusministerkonferenz vom 07.07.1972)“ (Beschluss der Kultusministerkonferenz vom 13.12.1973) in der jeweils geltenden Fassung,

die „Vereinbarung über Kenntnisse in Latein und Griechisch“ (Beschluss der Kultusministerkonferenz vom 26.10.1979) in der jeweils geltenden Fassung.

**I. Leistungen in den Kursen der Studienstufe****Kevin Baum**

Name

Fach	Bewertung Punktzahlen der einzelnen Kurse in einfacher Wertung			
	1. Halbjahr	2. Halbjahr	3. Halbjahr	4. Halbjahr

**Sprachlich-literarisch-künstlerisches Aufgabenfeld**

Deutsch	PF	10	12	13	13
Fremdsprachen (weitergeführt)					
Englisch		07	09	09	08
Fremdsprachen (neu aufgenommen)					
-.-.-.-		- - -	- - -	- - -	- - -
Darstellendes Spiel		13	14	- - -	- - -
-.-.-.-		- - -	- - -	- - -	- - -

**Gesellschaftswissenschaftliches Aufgabenfeld**

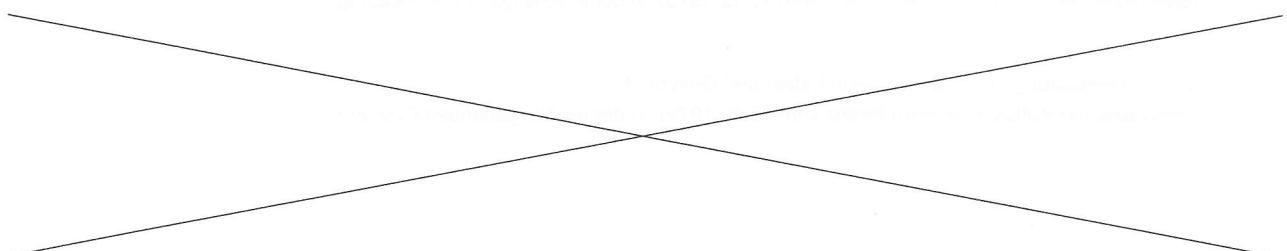
Gemeinschaftskunde		(12)	13	13	15
Geschichte	LF	12	12	13	13
Religion		(12)	13	14	(12)
-.-.-.-		- - -	- - -	- - -	- - -

**Mathematisch-naturwissenschaftlich-technisches Aufgabenfeld**

Mathematik	PF	13	14	14	15
Biologie		14	14	14	14
Chemie	LF	10	11	11	12
Informatik		(10)	13	- - -	- - -
-.-.-.-		- - -	- - -	- - -	- - -
Sport		(13)	(11)	(13)	(13)

(Leistungsfächer werden mit „LF“ gekennzeichnet. Die Bewertungen von Kursen, die nicht in die Gesamtqualifikation eingehen, sind in Klammern gesetzt.)

Die Beleg- und Einbringungsverpflichtung im Fach Deutsch / Fremdsprache / Mathematik wird durch den Kurs / die Kurse im Fach \_\_\_\_\_ erfüllt. 1)





**V. Bemerkungen****Kevin Baum**

Name

**keine Bemerkungen****VI. Herr****Kevin Baum**

hat mit der Ablegung der Abiturprüfung die Befähigung zum Studium an einer Hochschule in der Bundesrepublik Deutschland erworben.

Hamburg, 19.06.2006Prüfungsbeauftragte / Prüfungsbeauftragter

c. U. Pfleiderer

Schulleiterin / Schulleiter  
Abteilungsleiterin / Abteilungsleiter

Für die Umrechnung der Noten in Punkte gilt folgender Schlüssel:

Noten	sehr gut + 1 -	gut + 2 -	befriedigend + 3 -	ausreichend + 4 -	mangelhaft + 5 -	ungenügend 6
Punkte	15 14 13	12 11 10	9 8 7	6 5 4	3 2 1	0

# Universität des Saarlandes

Naturwissenschaftlich-Technische Fakultät I

## Urkunde Bachelor of Science

**Kevin Christian Helmut Werner  
Baum**

geboren am 28.07.1986 in Schwetzingen,

hat die Prüfung zum Bachelor of Science (Informatik) gemäß der  
Prüfungsordnung vom 8. Juni 2006 bestanden.

Ihm wird hiermit der akademische Grad

**Bachelor of Science (B.Sc.)**

verliehen.

Saarbrücken, den 24.11.2011

Der Dekan  
Naturwissenschaftlich-Technische Fakultät I

Univ.-Prof. Dr.-Ing. H. Hermanns

Der Vorsitzende  
Prüfungsausschuss Informatik

Univ.-Prof. B. Finkbeiner, Ph.D.





# Veranstaltungsübersicht / Zeugnis

für

## Kevin Christian Helmut Werner Baum

geboren am 28.07.1986 in Schwetzingen

Semester	Titel	Prüfer	LP	Note
SS 2011	Quadrupoles in PDE-based Image Compression using Stochastic Masks (Bachelorseminar)	Prof. Dr. J. Weickert	9 LP	1,7
WS 2007/2008	Programmierung 1	Prof. Dr. H. Hermanns	9 LP	2,0
SS 2008	Programmierung 2	Prof. Dr. S. Hack	9 LP	2,0
SS 2008	Systemarchitektur	Prof. Dr. W. Paul	9 LP	2,7
WS 2008/2009	Mathematik für Informatiker 3	Prof. Dr. V. John	9 LP	1,7
WS 2007/2008	* Analysis 1	Prof. Dr. R. Schulze-Pillot-Ziemer	9 LP	2,0
SS 2008	Analysis 2	Prof. Dr. R. Schulze-Pillot-Ziemer	9 LP	3,7
WS 2007/2008	* Lineare Algebra 1	Prof. Dr. F.-O. Schreyer	9 LP	2,0
SS 2008	Lineare Algebra 2	Prof. Dr. F.-O. Schreyer	9 LP	1,7
SS 2010	Emerging Methods for Image Compression (Seminar)	PD Dr. M. Breuss	7 LP	1,7
WS 2008/2009	Grundzüge der Theoretischen Informatik	Prof. Dr. M. Bläser	9 LP	1,3
WS 2008/2009	Grundzüge von Algorithmen und Datenstrukturen	Prof. Dr. R. Seidel	6 LP	1,0
WS 2009/2010	Image Processing and Computer Vision	Prof. Dr. J. Weickert	9 LP	2,3
SS 2009	Informationssysteme	Prof. Dr. J. Dittrich	6 LP	1,7
SS 2009	Introduction to Computational Logic	Prof. Dr. G. Smolka	9 LP	3,7
SS 2009	Nebenläufige Programmierung	Prof. Dr. H. Hermanns	6 LP	2,7
SS 2009	Probleme in der diskreten kombinatorischen Geometrie	Prof. Dr. R. Seidel	5 LP	2,0
SS 2010	Informationssysteme (Tutor)	Prof. Dr. J. Dittrich	4 LP	mit Erfolg teilgenommen
WS 2007/2008	Ringvorlesung über Perspektiven der Informatik	Prof. Dr. R. Wilhelm	4 LP	mit Erfolg teilgenommen

WS 2008/2009	Softwarepraktikum	Prof. Dr. A. Zeller	9 LP	mit Erfolg teilgenommen
WS 2008/2009	C/C++	Dr. P. Lucas	2 LP	mit Erfolg teilgenommen
SS 2010	Differential Equations in Image Processing and Computer Vision	Dr. A. Bruhn	9 LP	mit Erfolg teilgenommen
WS 2008/2009	Unix	Prof. Dr. S. Hack	2 LP	mit Erfolg teilgenommen

\* Nebenfach: Mathematik

Bachelorarbeit	Prüfer	LP	Note
Stützstellenauswahl für diffusionsbasierte Bildkompression unter Berücksichtigung einer Quadixel-Substruktur-Restriktion	Prof. Dr. J. Weickert Dr. A. Bruhn	12	1,3

Datum der letzten Prüfung: 16.08.2011

Erhaltene Leistungspunkte: 180

## Kevin Christian Helmut Werner Baum

hat den Bachelor of Science Studiengang Informatik mit der Gesamtnote

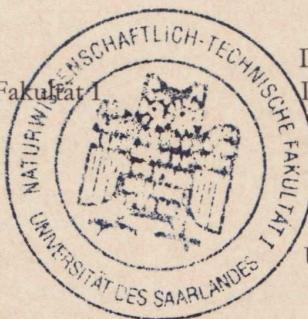
**gut (2,0)**

abgeschlossen.

Saarbrücken, den 24.11.2011

Der Dekan  
Naturwissenschaftlich-Technische Fakultät I

Univ.-Prof. Dr.-Ing. H. Hermanns



Der Vorsitzende  
Prüfungsausschuss Informatik

Univ.-Prof. B. Finkbeiner, Ph.D.

# Universität des Saarlandes

Naturwissenschaftlich-Technische Fakultät I

## Urkunde Master of Science

**Kevin Christian Helmut Werner  
Baum**

geboren am 28.07.1986 in Schwetzingen,

hat die Prüfung zum Master of Science (Informatik) gemäß der  
Prüfungsordnung vom 8. Juni 2006 bestanden.

Ihm wird hiermit der akademische Grad

**Master of Science (M.Sc.)**

verliehen.

Saarbrücken, den 22.08.2013

Der Dekan  
Naturwissenschaftlich-Technische Fakultät I

*M. D. Groves*

Univ.-Prof. Dr. M. D. Groves

Der Vorsitzende  
Prüfungsausschuss Informatik

*B. Finkbeiner*

Univ.-Prof. B. Finkbeiner, Ph.D.





# Veranstaltungsübersicht / Zeugnis

für

## Kevin Christian Helmut Werner Baum

geboren am 28.07.1986 in Schwetzingen

Semester	Titel	Prüfer	LP	Note
WS 2009/2010	Computer Graphics	Prof. Dr. P. Slusallek	9 LP	1,7
WS 2010/2011	Correspondence Problems in Computer Vision	Prof. Dr. A. Bruhn	6 LP	1,0
WS 2010/2011	Einführung in die Sprachphilosophie und Logik	Prof. Dr. N. Strobach	9 LP	1,0
WS 2011/2012	Image Compression	Prof. Dr. J. Weickert	6 LP	1,3
SS 2012	Kryptographie	Prof. Dr. M. Backes	9 LP	2,3
WS 2011/2012	Probabilistic Graphical Models and their Applications	Prof. Dr. B. Schiele	6 LP	1,0
WS 2010/2011	Recent Advances in Image Processing and Computer Vision (Seminar)	PD Dr. M. Breuß	7 LP	1,3
WS 2010/2011	Security	Prof. Dr. M. Backes/ Dr. M. Maffei	9 LP	1,0
SS 2012	Parallel PDE-based Image Decompression-Speeding Things up Using CUDA (Masterseminar)	Prof. Dr. J. Weickert	12 LP	1,3
SS 2012	Automated Reasoning	Prof. Dr. C. Weidenbach	9 LP	mit Erfolg teilgenommen
WS 2010/2011	Einführung in die Ethik	Prof. Dr. U. Wessels	9 LP	mit Erfolg teilgenommen

Masterarbeit	Prüfer	LP	Note
GPGPU-gestützte diffusionsbasierte naive Videokompression	Prof. Dr. J. Weickert Prof. Dr. T. Herfet	30	1,7

Datum der letzten Prüfung: 15.01.2013

Erhaltene Leistungspunkte: 121

# Kevin Christian Helmut Werner Baum

hat den Master of Science Studiengang Informatik mit der Gesamtnote

**sehr gut (1,4)**

abgeschlossen.

Saarbrücken, den 22.08.2013

Der Dekan  
Naturwissenschaftlich-Technische Fakultät

*M. D. Groves*

Univ.-Prof. Dr. M. D. Groves

Der Vorsitzende  
Prüfungsausschuss Informatik

*B. Finkbeiner*

Univ.-Prof. B. Finkbeiner, Ph.D.



# UNIVERSITÄT DES SAARLANDES

Philosophische Fakultät I

Geschichts- und Kulturwissenschaften

verleiht unter dem Dekanat von  
Herrn Univ.-Prof. Dr. phil. Peter Riemer

mit dieser Urkunde

**Herrn**

**Kevin Christian Helmut Werner Baum**

geboren am 28. Juli 1986 in Schwetzingen

den akademischen Grad

**Master of Arts (M.A.)**

im Master-Studiengang

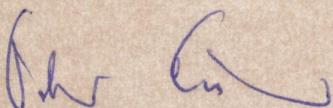
**Philosophie**

mit der Gesamtnote

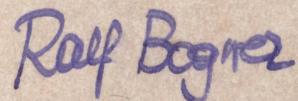
**SEHR GUT (1,1)**

Tag der letzten Prüfungsleistung war der 5. Juni 2014.

Saarbrücken, 29. August 2014



Der Dekan der  
Philosophischen Fakultät I  
(Univ.-Prof. Dr. P. Riemer)



Der Vorsitzende  
des Prüfungsausschusses  
(Univ.-Prof. Dr. R. Bogner)

## Übersicht über alle Leistungen

Seite 1 von 3

Name des Studierenden: Kevin Christian Helmut Werner Baum  
 Geburtsdatum und -ort: 28.07.1986 in Schwetzingen  
 (angestrebter) Abschluss: Master (Kernbereich)  
 Studiengang: Philosophie  
 Fachsemester: 5  
 Matrikelnummer: 2518184  
 Heimathochschule: Universität des Saarlandes

Prüfungsnr	Bezeichnung der Leistung	Semester	Prüf. Datum	Note	Status	Bonus	Vermerk	Versuch
1010	<b>Gesamtkonto</b>						<b>99</b>	
1050	<b>Pflichtmodule</b>			1,0			<b>36</b>	
1320	<b>MS-PP: Praktische Philosophie</b>			1,0			<b>18</b>	
1321	<b>MS-PP1: Praktische Philosophie in Gegenwart und Geschichte (MS-9CP) / MS-PP2: Praktische Philosophie in Gegenwart und Geschichte (MS-9CP)</b>			1,0			<b>18</b>	
10081	MS-PP1: Praktische Philosophie in Gegenwart und Geschichte (MS-9CP) / MS-PP2: Praktische Philosophie in Gegenwart und Geschichte (MS-9CP)	WiSe12/13		1,0	bestanden	18		1
1330	<b>MS-TP: Theoretische Philosophie</b>			1,0			<b>18</b>	
1331	<b>MS-TP1: Theoretische Philosophie (MS-9CP) / MS-TP2: Theoretische Philosophie (MS-9CP)</b>			1,0			<b>18</b>	
10090	MS-TP1: Theoretische Philosophie (MS-9CP) / MS-TP2: Theoretische Philosophie (MS-9CP)	WiSe12/13		1,0	bestanden	18		1

Falls erforderlich, Liste auf getrenntem Blatt fortsetzen

Hinweis: Module, deren Teilleistungen alle bestanden werden müssen, können Teilleistungen mit 0 Credit-Points enthalten. Diese Credit-Points werden direkt auf das Modulelement gebucht, wenn alle Teilleistungen bestanden sind!

# Übersicht über alle Leistungen

Seite 2 von 3

Kevin Christian Helmut Werner Baum

PrüfungsNr	Bezeichnung der Leistung	Semester	Prüf. Datum	Note	Status	Bonus	Vermerk	Versuch
1070	<b>Wahlbereich</b>			1,0			<b>27</b>	
900	<b>Dummymodul 1</b>			1,0			<b>13,5</b>	
910	<b>Einführung in die Wissenschaftstheorie</b>			1,7			<b>4,5</b>	
950	Einführung in die Wissenschaftstheorie	WiSe11/12		1,7 bestanden		4,5		1
911	<b>Geschichte der Philosophie Antike/Neuzeit</b>			1,0			<b>9</b>	
951	Geschichte der Philosophie Antike/Neuzeit	SoSe12		1,0 bestanden		9		1
901	<b>Dummymodul 2</b>			1,0			<b>13,5</b>	
912	<b>Security (Informatik)</b>			1,0			<b>9</b>	
952	Security (Informatik)	SoSe13		1,0 bestanden		9		1
913	<b>Einführung in die Erkenntnistheorie</b>			1,0			<b>4,5</b>	
953	Einführung in die Erkenntnistheorie	WiSe12/13		1,0 bestanden		4,5		1
1060	<b>Wahlflichtmodule</b>			1,0			<b>36</b>	
1350	<b>MS-E: Ergänzungsmodul</b>			1,0			<b>18</b>	
1351	<b>Tempus &amp; Modalität</b>			1,0			<b>18</b>	
10096	Tempus & Modalität	WiSe12/13		1,0 bestanden		18		1
1340	<b>MS-VT: Vertiefungsmodul</b>			1,0			<b>18</b>	
1341	<b>MS-VT1: Vertiefungsmodul element 1 (MS-9CP) / MS-VT2: Vertiefungsmodul element 2 (MS-9CP)</b>			1,0			<b>18</b>	
10091	Antinaturalismus	WiSe13/14		1,0 bestanden		18		1
<b>Gesamtnote</b>				1,0				

Falls erforderlich, Liste auf getrenntem Blatt fortsetzen

Hinweis: Module, deren Teilleistungen alle bestanden werden müssen, können Teilleistungen mit 0 Credit-Points enthalten. Diese Credit-Points werden direkt auf das Modulelement gebucht, wenn alle Teilleistungen bestanden sind!

## Übersicht über alle Leistungen

### Erläuterungen

#### (1) Beschreibung des Notensystems, das an der Hochschule angewendet wird

1,0/ 1,3	sehr gut
1,7/ 2,0/ 2,3	gut
2,7/ 3,0/ 3,3	befriedigend
3,7/ 4,0	genügend
5,0	ungenügend

**Promotionsausschuss**

Technische Universität Dortmund | D-44227 Dortmund / Dekanat

Herrn  
Kevin Baum  
Martin-Luther-Straße 28  
66111 Saarbrücken  
per Mail: [mail@kevinbaum.de](mailto:mail@kevinbaum.de)

Prof. Dr. M. Basse / Vorsitzender  
Emil-Figge-Str. 50  
D-44227 Dortmund  
T 0231.755.2866  
T 0231.755.2806 (He)  
[maria.hemker@tu-dortmund.de](mailto:maria.hemker@tu-dortmund.de)  
[michael.basse@tu-dortmund.de](mailto:michael.basse@tu-dortmund.de)  
[www.tu-dortmund.de](http://www.tu-dortmund.de)

Diktatzeichen                  Aktenzeichen                  Ort                  Datum                  Dienstgebäude/Raum  
Ba/He                                   Dortmund                  25.03.2024                  EF 50 / 0.435

---

**B E S C H E I N I G U N G**

Hiermit wird Herrn Kevin C. H. W. Baum, geboren am 28.07.1986 in Schwetzingen,  
bescheinigt, dass er am 22.03.2024 erfolgreich promoviert wurde.

Die Dissertation

**Doing Wrong with Others**  
**Multi-Agent Consequentialism as a Solution for the Collective Action Problem**

wurde mit der Note **summa cum laude** bewertet.

Note der Disputation: **summa cum laude**

Als Prädikat für die Promotion wurde **summa cum laude** festgelegt.

Auflagen zur Veröffentlichung der Dissertation

- keine  
 folgende:

Gemäß §18 PromO ist die Dissertation in angemessener Weise der wissenschaftlichen Öffentlichkeit durch Vervielfältigung und Verbreitung zugänglich zu machen. Die Dissertation ist spätestens ein Jahr nach der mündlichen Prüfung zu veröffentlichen. (In begründeten Ausnahmefällen kann der Vorsitzende des Promotionsausschusses die Frist verlängern.)

Bitte beachten Sie die Möglichkeiten der Veröffentlichung lt. der entsprechenden PromO. Gleichzeitig mache ich Sie auf die Regelungen zur Abgabe von Veröffentlichungen der UB der Technischen Universität aufmerksam:

(Deutsch: [https://www.ub.tu-dortmund.de/Eldorado/abgabe\\_dissertationen.html](https://www.ub.tu-dortmund.de/Eldorado/abgabe_dissertationen.html) /  
Englisch: [http://www.ub.tu-dortmund.de/Eldorado/abgabe\\_dissertationen.html.en](http://www.ub.tu-dortmund.de/Eldorado/abgabe_dissertationen.html.en)).

Ich verweise nach §19 PromO daraufhin, dass erst mit Aushändigung der Promotionsurkunde das Recht zur Führung des Doktorgrades entsteht. Die Führung des Titels „Dr. des.(ignatus)“ ist nicht zulässig.

Freundliche Grüße



Prof. Dr. Michael Basse

## **Research Publications**

---

 → That symbol indicates items in this CV and elsewhere in this application which I propose to be cumulatively equivalent to a habilitation; see also the separate document »Habilitation Equivalence«.

### **Journal Articles**

-  S. Biewer, K. Baum, S. Sterz, *et al.*, "Software Doping Analysis for Human Oversight," *Formal Methods in System Design*, 2024.  doi: 10.1007/s10703-024-00445-2.
-  K. Baum, J. Bryson, F. Dignum, *et al.*, "From Fear to Action: AI Governance and Opportunities for All," *Frontiers in Computer Science*, vol. 5, 2023.  doi: 10.3389/fcomp.2023.1210421.
-  K. Baum and S. Sterz, "Ethics for Nerds," *The International Review of Information Ethics*, vol. 31, no. 1, 2022, special issue on »Ethics in the Age of Smart Systems«.  doi: 10.29173/irie484.
-  K. Baum, S. Mantel, E. Schmidt, and T. Speith, "From Responsibility to Reason-Giving Explainable Artificial Intelligence," *Philosophy & Technology*, vol. 35, no. 1, 2022.  doi: 10.1007/s13347-022-00510-w.
-  N. Schlicker, M. Langer, S. K. Ötting, K. Baum, C. J. König, and D. Wallach, "What to Expect from Opening Up 'Black Boxes'? Comparing Perceptions of Justice Between Human and Automated Agents," *Computers in Human Behavior*, vol. 122, 2021.  doi: 10.1016/j.chb.2021.106837.
-  M. Langer, D. Oster, T. Speith, *et al.*, "What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research," *Artificial Intelligence*, vol. 296, 2021.  doi: 10.1016/j.artint.2021.103473.
-  M. Langer, K. Baum, C. J. König, V. Hähne, D. Oster, and T. Speith, "Spare Me the Details: How the Type of Information About Automated Interviews Influences Applicant Reactions," *International Journal of Selection and Assessment*, vol. 29, no. 2, pp. 154–169, 2021.  doi: 10.1111/ijsa.12325.

### **Conference Proceedings**

-  D. Baum, K. Baum, T. P. Gros, and V. Wolf, "XAI Requirements in Smart Production Processes: A Case Study," in *Explainable Artificial Intelligence. Proceedings of the World Conference on eXplainable Artificial Intelligence (xAI 2023)*, L. Longo, Ed., ser. Communications in Computer and Information Science (CCIS), vol. 1901, Cham: Springer Nature Switzerland, 2023, pp. 3–24.  doi: 10.1007/978-3-031-44064-9\_1.
-  M. Langer, K. Baum, K. Hartmann, S. Hessel, T. Speith, and J. Wahl, "Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives," in *29th IEEE International Requirements Engineering Conference Workshops (RE 2021 Workshops), Notre Dame, Indiana, USA*, T. Yue and M. Mirakhorli, Eds., IEEE, 2021, pp. 164–168.  doi: 10.1109/REW53955.2021.00030.
-  S. Sterz, K. Baum, A. Lauber-Rönsberg, and H. Hermanns, "Towards perspicuity requirements," in *29th IEEE International Requirements Engineering Conference Workshops (RE 2021 Workshops), Notre Dame, Indiana, USA*, T. Yue and M. Mirakhorli, Eds., IEEE, Sep. 2021, pp. 159–163.  doi: 10.1109/REW53955.2021.00029.

- 4** M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, "Explainability as a Non-Functional Requirement," in *27th IEEE International Requirements Engineering Conference (RE 2019), Jeju Island, South Korea*, IEEE, 2019, pp. 363–368.  DOI: 10.1109/RE.2019.00046.
- 5** K. Baum, H. Hermanns, and T. Speith, "Towards a framework combining machine ethics and machine explainability," in *Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology (CREST 2018), Thessaloniki, Greece, 21st April 2018*, B. Finkbeiner and S. Kleinberg, Eds., 2019.  DOI: 10.4204/EPTCS.286.4.
- 6** K. Baum, N. Kirsch, K. Reese, P. Schmidt, L. Wachter, and V. Wolf, "Informatikunterricht in der Grundschule? Erprobung und Auswertung eines Unterrichtsmoduls mit Calliope mini," in *Proceedings of the Informatik für alle, 18. GI-Fachtagung Informatik und Schule (INFOS 2019) in the GI-Edition: Lecture Notes in Informatics (LNI)*, A. Pasternak, Ed., vol. P-288, Gesellschaft für Informatik, 2019, pp. 49–58.  DOI: 10.18420/INFOS2019-B1.
- 7** A. Sesing and K. Baum, "Anforderungen an die Erklärbarkeit maschinengestützter Entscheidungen," in *Die Macht der Daten und der Algorithmen – Regulierung von IT, IoT und KI. Tagungsband DSRI-Herbstakademie 2019*, J. Taeger, Ed., 2019, pp. 435–449.  URL: <http://olwir.de/?content=reihen/uebersicht&sort=tb&isbn=978-3-95599-061-9>.
- 8** K. Baum, H. Hermanns, and T. Speith, "From Machine Ethics To Machine Explainability and Back," in *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2018), Fort Lauderdale, Florida, USA*, 2018.  URL: [https://isaim2018.cs.ou.edu/papers/ISAIM2018\\_Ethics\\_Baum\\_etal.pdf](https://isaim2018.cs.ou.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf).
- 9** K. Baum, M. Köhl, and E. Schmidt, "Two Challenges for CI Trustworthiness and How to Address Them," in *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*, M. Pereira-Fariña and C. Reed, Eds., Dundee, United Kingdom: Association for Computational Linguistics, Sep. 2017.  DOI: 10.18653/v1/W17-3701.
- 10** K. Baum, "What the Hack Is Wrong with Software Doping?" In *Proceedings of the 7th International Symposium on Leveraging Applications of Formal Methods: ISoLA 2016: Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications*, T. Margaria and B. Steffen, Eds., ser. Lecture Notes in Computer Science (LNCS), Springer International Publishing, vol. 9953, 2016, pp. 633–647.  DOI: 10.1007/978-3-319-47169-3\_49.
- 11** A. Krekhov, J. Grüninger, K. Baum, D. McCann, and J. Krüger, "MorphableUI: A Hypergraph-Based Approach to Distributed Multimodal Interaction for Rapid Prototyping and Changing Environments," in *Proceedings of the 24th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2016) in co-operation with EUROGRAPHICS*, ser. Computer Science Research Notes (CSRN), Václav Skala-UNION Agency, 2016, pp. 299–308.  URL: [http://wscg.zcu.cz/WSCG2016/!!\\_CSRN-2602.pdf](http://wscg.zcu.cz/WSCG2016/!!_CSRN-2602.pdf).

## Chapters

- 1** K. Baum, "Utilitarismus und das Problem kollektiven Handelns," in *Handbuch Utilitarismus*, V. Andrić and B. Gesang, Eds., *in print*, J.B. Metzler.

## Submitted and Under Review

- 1 K. Baum, S. Biewer, H. Hermanns, et al., *Taming the AI Monster: Monitoring of Individual Fairness for Effective Human Oversight*, submitted as invited paper for the proceeding of the 30th International SPIN symposium on Model Checking of Software (SPIN 2024) colocated with the 27th European Joint Conferences on Theory and Practice of Software (ETAPS 2024), thoroughly revised version (with changed focus) of Biewer et al. 2024, 2024.
- 2 S. Sterz, K. Baum, S. Biewer, et al., *On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives*, submitted to and under review at the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 2024.
- 3 L. M. Budde and K. Baum, *Building Conceptual Bridges: On the Non-Functional Analysis of AI-Systems*, submitted to the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 2024.
- 4 K. Baum and L. Dargasz, *A Reason-Responsiveness Approach to Machine Ethics for Reinforcement-Learning Agents*, extended abstract, submitted to the international conference on Formal Ethics 2024 at the University of Greifswald, Greifswald, Germany, 2024. Ⓢ URL: <https://www.wiko-greifswald.de/formal-ethics-2024/>.
- 5 K. Baum, *A Philosophical Foundation for Machine Ethics: A Pluralism in Decision Procedures*, extended abstract, submitted to the international conference on Formal Ethics 2024 at the University of Greifswald, Greifswald, Germany, 2024. Ⓢ URL: <https://www.wiko-greifswald.de/formal-ethics-2024/>.
- 6 N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer, *A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings of the Trustworthiness Assessment Model (TrAM)*, submitted to and under review at Computers in Human Behavior for the special issue on *The Social Bridge: An Interdisciplinary View on Trust in Technology*, 2023. Ⓢ DOI: 10.31234/osf.io/qhwvx.
- 7 M. Langer, K. Baum, and N. Schlicker, *Effective Human Oversight of AI-based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs*, submitted to and under second review after minor revision at *Minds and Machines* for the special issue on *Interdisciplinary Perspectives on the (Un)fairness of Artificial Intelligence*, 2023. Ⓢ DOI: 10.31234/osf.io/ke256.

## Publications Aimed at Public Audience



### Publications (Selection)

- 1 K. Baum and S. Sterz, "Ethics for Nerds: Verantwortung für die Informatiker:innen von Morgen," *Mitteilungen des Fachverbandes Philosophie*, vol. 63, pp. 72–83, 2023. Ⓢ URL: <https://www.fv-philosophie.de/wp-content/uploads/2024/02/Mitteilungen-2023-Hompage-mit-kommentiertem-Teilabdruck-der-Beitraege.pdf>.
- 2 K. Baum, "Sicherheitsrisiko KI," *IMPULSE – Das Magazin der VolkswagenStiftung*, pp. 38–39, 2022. Ⓢ URL: <https://www.volkswarenstiftung.de/de/news/aktuelles/sicherheitsrisiko-ki>.
- 3 K. Baum, "Moral für Maschinen?" *OPUS Kulturmagazin*, no. 85, 2021. Ⓢ URL: <https://opus-kulturmagazin.de/moral-fuer-maschinen/>.

- 4 M. Schillo and K. Baum, "»Have you ever questioned the nature of your reality?« — Westworld aus der Perspektive der Philosophie," in *Folge um Folge. Multiple Perspektiven auf die Fernsehserie*, Universitätsverlag Hildesheim, 2020, ch. 13.  doi: 10.18442/160.
- 5 K. Baum, T. Feldkamp, S. Henning, and N. Käfer, "Algorithmen und die Grenzen von Fairness," in *Unternehmensverantwortung im digitalen Wandel*, Bertelsmann Stiftung and Wittenberg-Zentrum für Globale Ethik, Ed., Verlag Bertelsmann Stiftung, 2019, ch. 4.  doi: 10.11586/2020063.
- 6 M. Langer, K. Baum, and C. J. König, "Die (un-)nachvollziehbarkeit algorithmenbasierter entscheidungen: Implikationen und empfehlungen für die zukunft," in *Mensch und Gesellschaft im digitalen Wandel*, Deutscher Psychologen Verlag, 2018, pp. 32–38, ISBN: 978-3-942761-47-5.  doi: 10.22028/D291-31275.
- 7 M. Langer, K. Baum, and C. J. König, "Algorithmen bei der Personalauswahl – eine kritische und hoffnungsvolle Betrachtung," *Wirtschaftspsychologie aktuell*, 1st ser., vol. 25, pp. 36–42, 2018, ISSN: 1611-9207.  doi: 10.22028/D291-31272.

## **Third-Party Funding**

---

### **Current Projects**

---

Altogether, my direct or indirect responsibilities encompass managing project funding that totals at least **700,000 euros**. Below is a detailed breakdown.

#### **Health.AI – Ethical and Legal Implications (Health.AI ELI)** – running since January 2024

*Link:* <https://health-ai.de/health-ai-eli/>

*Context & Core Idea:* Research project on normative questions accompanying the *Health.AI network* (»normative Begleitforschung«), a BMBF WIR!-network in Saarland (»WIR!« stands for »Wandel durch Innovation in der Region«).

*My Role:* Initiator, provider of the original idea, driving force, and lead proposal writer (for Algoright). The relevant share of funds was awarded to Algoright, and the position is currently filled by Thorsten Helfer (an ethicist at the UdS philosophy department).

*Funding:* Total funding amount **approx. 700,000 euros**, of which **approx. 230,000 euros** for basic research in the field of applied ethics by Algoright e.V. (30 person-months (PM) plus funds for consulting subcontractors).

*Reference:* Thorsten Helfer, [thorsten.helfer@algoright.de](mailto:thorsten.helfer@algoright.de)

#### **Towards a Centre for European Research in Trusted Artificial Intelligence (ToCERTAIN)**

– running since June 2023

*Link:* <https://certain.dfk.de>

*Context & Core Idea:* The ToCERTAIN project is funded as part of the Saarland funding for research and infrastructure at universities and non-university research institutions, which in turn is funded by the *European Regional Development Fund* (ERDF, aka. EFRE). ToCERTAIN lays the foundation of the *Centre for European Research in Trusted Artificial Intelligence* (CERTAIN), which aims to bring trusted artificial intelligence into practice in close and intensive cooperation with partners from research and industry – in Saarland, the Greater Region and throughout Europe.

*My Role:* I have been head of the ToCertain project and of the center since December 2023. Since then, vacancies have been filled, the strategic course of action was defined, and research priorities were set under my responsibility. Besides that, in my role as deputy head and research area manager of NMM, I am responsible for more than a quarter of the project funds.

*Funding:* The total funding amounts to approx. **2 million euros** over three years, acquired by the research department for *Agents and Simulated Reality* (ASR) of Professor Philipp Slusallek. As of April 1, the project structure has been spread across several research departments, including *Neuro-Mechanistic Modeling* (NMM). The share of NMM is **approx. 500,000 euros**.

*Reference:* Prof. Verena Wolf, [verena.wolf@dfki.de](mailto:verena.wolf@dfki.de)

#### **Foundations of Perspicuous Software Systems/Center for Perspicuous Computing (CPEC)**

– running since 2020

*Link:* <https://www.perspicuous-computing.science/>

*Context & Core Idea:* The *Center for Perspicuous Computing* (CPEC) is a DFG *Transregional Collaborative Research Centre* (CRC 248). CPEC is a cooperation of *Saarland University*, *Technische Universität Dresden*, *Max Planck Institute for Software Systems*, and *CISPA Helmholtz Center for Information Security* jointly with experts from *TU Wien*. I am mainly involved in the subproject *E7 – Perspicuity and Societal Risk*.

*My Role:* I was an active contributor to the project proposal, especially concerning subproject E7 – *Perspicuity and Societal Risk*, which primarily investigates normative questions connected to perspicuity, especially the interplay of technical perspicuity properties (from transparency to explainability to traceability) in connection with the desideratum of effective human oversight in the sense of, for example, Article 14 of the EU AI Act. I remain an active affiliated project researcher after moving to DFKI on January 1, 2023. I was given the prospect of becoming a *member* of the CPEC in September so that I could then officially become a PI with my own human resources in E7. There is the prospect of becoming PI for ethics of digitalization in a third funding round.

*Funding:* The current second funding round of CPEC is funded with around **13 million euros**. The E7 sub-project receives over **700,000 euros**.

*Reference:* CPEC speaker Prof. Holger Hermanns, [hermanns@cs.uni-saarland.de](mailto:hermanns@cs.uni-saarland.de)

### **Explainable Intelligent Systems (EIS)**

– running since 2018

*Link:* <https://www.eis.science/>

*Context & Core Idea:* EIS is a highly interdisciplinary project of computer science, philosophy, psychology, and law researching the connection between explainability requirements and social desiderata. What began as a project at the Uds is now a collaboration between the Uds (Prof. Holger Hermanns, Prof. George Borges, and Prof. Ulla Wessels), the University of Bayreuth (Prof. Lena Kästner), TU Dortmund University (Prof. Eva Schmidt), and the University of Freiburg (Prof. Markus Langer). Officially expires in Q2 2024, partially extended cost-neutrally.

*My Role:* I am the source of the original idea, the driving force of and an active contributor to the proposals for the preliminary and main funding phase. Later, I was a research assistant in the project for both computer science and philosophy and the project coordinator for a while. I remain an active affiliated project member after moving to DFKI on January 1, 2023.

*Funding:* EIS was initially funded with **150,000 euros** for a planning phase (2018–2019) and then with **1.5 million euros** in the second project phase (2020–2024) by *VolkswagenStiftung* within the track *Artificial Intelligence and the Society of the Future*.

*References:* EIS speaker Prof. Lena Kästner, [Lena.Kaestner@uni-bayreuth.de](mailto:Lena.Kaestner@uni-bayreuth.de)

### **Proposals under Review or Recently (Pre-)Approved**

---

In the last few months, I initiated further proposals that are now under review or have recently been positively evaluated. In the event of success, I would have (partly shared) access to roughly another **approx. 700,000 euros**. Here are the details.

#### **Multi-Level Abstractions and Causal Modeling for Enhanced Reinforcement Learning**

(**MAC-MERLin**) – starting July or August 2024

*Context & Core Idea:* MAC-MERLin is a project within the context of the DFKI »Rahmenvertragsmittel« of the *German Federal Ministry of Education and Research* (BMBF) with the DFKI. This 36-month project will shape the fundamental direction of the NMM research department at DFKI. We will explore the possible combination of different levels of abstraction in problem representation and causal modeling with deep-learning-based reinforcement learning for practical applications.

*My Role:* Co-author and principal investigator of work package 4 (WP4) on explainability and mechanistic interpretability, emphasizing topics like trust calibration and effective human oversight.

*Funding:* The overall project has a volume of **1.8 million euros**, WP4 comprises 27 PM, **approx. 330,000 euros** plus travel and some minor »Sachmittel«.

*Chance of Success:* The project has been approved by the DFKI Scientific Advisory Board on February 29, 2024. The formal application for BMBF (via AZK) is now being prepared and will be submitted at the end of March. Funding is expected to be available from July or August 2024.

*Reference:* Prof. Verena Wolf, [verena.wolf@dfki.de](mailto:verena.wolf@dfki.de)

## **DATIpilot Innovation Community Pink Box**

- submitted October 2023

*Context & Core Idea:* Proposal for a network project called »Soziale Innovationscommunity« in context of the DATIpilot call of the BMBF, where »DATI« stands for »Deutsche Agentur für Transfer und Innovation«. *Pink Box* is designed as a bidirectional interface between science and civil society. Societal and sustainability-related challenges (in the sense of the *UN Sustainable Development Goals*, SDGs) are communicated through participatory science communication offerings characterized by creativity and art. At the same time, idea workshops and hackathons are used to realize concrete application ideas for AI in everyday community life as part of so-called *community projects*.

*My Role:* Principal investigator, provider of the original idea, and the person who brought the consortium together. I am the main author by name for both DFKI and Algoright e.V.

*Funding:* Total funding volume of slightly over **1 million euros**, DFKI's share is **approx. 210,000 euros** and Algoright's share is another **180,000 euros**. In addition, spin-off projects are possible and easy to implement in the form of so-called *community projects* (»Community-Projekte«).

*Chance of Success:* Medium, shortlist decision expected at the end of March.

*Reference:* Anna Lawera (CEO of East Side Fab e.V.), [a.lawera@eastsidefab.de](mailto:a.lawera@eastsidefab.de)

Further applications are in various stages of conception or preparation (see *Research Plan* attached for more details).

The written applications of most of the projects listed here can be made available to the selection committee on request.

## **Research Plan**

---

Here are my research priorities for the upcoming years, including concrete projects in various stages of planning and execution.

### **General Research Directions**

---

My research projects for the next years can be clustered in terms of four research directions, which I would like to tackle within different time frames.

#### **Research Direction A: *Effective/meaningful Human Oversight*** (short to mid-term)

Effective (or meaningful) human oversight is a central element of ethical guidelines and global legislation aimed at regulating the use of AI-based systems, especially in high-risk situations. Various laws, including the European AI Act, mandate effective human oversight as a fundamental tool for risk mitigation. Perspicuity requirements, above all explainability, are plausibly a central prerequisite for human overseers to be more than ethical fig leaves and obedient button-pushers. For this reason, my research in the context of the *Explainable Intelligent Systems* (EIS) and the CRC Center for *Perspicuous Computing* (CPEC) has increasingly focused on the conceptual and normative conditions for effective human oversight.

I would like to broaden and deepen this research, specifically concerning perspicuity as a plausible prerequisite for successful and ethical hybrid AI, i.e., the combination of human and artificial intelligence. The focus will lie on questions of effective human oversight, accountability, including (moral) responsibility, justified and calibrated trust in AI, and, more generally, ethically robust hybrid decision-making. I will also focus on the purposeful advancement of existing XAI methods and address questions of verifiability and certifiability of the fulfillment of those desiderata.

#### **Research Direction B: *Ethical Quality Measures for Models*** (mid-term)

Even though there is a plethora of attributes associated with *Trustworthy AI*, *Trusted AI*, and *Responsible AI*, including explainability, fairness, privacy protection, and human autonomy, it remains vague what makes one model (ethically speaking) better than another in an overall sense. Neither the relationships nor the dependencies between these quality measures, including (necessary?) trade-offs and the impact of context on these aspects, are well understood.

My research will explore these relationships to enhance our comprehension of ethical AI models. My working hypothesis is that there is no universal hierarchy between the plurality of quality measures but that their validity depends on the application domain and the embedding in larger decision-making processes. My research aims to establish ethically well-founded holistic best practices and design guidelines and to derive concrete regulatory recommendations.

#### **Research Direction C: *Moral-Philosophically Informed Machine Ethics*** (mid to long-term)

Machine ethics explores embedding moral principles into the decision-making of AI systems. This field intersects normative and applied ethics as well as computer science, incorporating normative ethics, safety engineering, and the fields of verification and validation. Current research programs in machine ethics as a subfield of informatics include attempts to translate complex ethical theories into algorithms without considering the distinction between criteria of rightness and decision procedures; implementing highly specific rule sets tailored to particular domains without theoretical foundations; and integrating 'moral' learning signals into reinforcement learning processes overlooking the relationships of different normative domains. Purely philosophical debates, in contrast, often miss the practical goal of machine ethics and ignore the technological sophistication of AI agents as well as insights from machine learning and safety research.

My goal is to establish a robust theoretical foundation for machine ethics. In addition to creating a moral-philosophical background theory for machine ethics, I aim to bridge the gap between disciplines through interdisciplinary conferences and projects.

**Research Direction D: A Philosophy of Society-Scale Feedback Loops**  
**(mid to long-term)**

AI systems significantly influence human behavior through mechanisms like profiling and nudging and are influenced by human behavior via the data this behavior generates. However, beyond this individual feedback loop, there exists a larger feedback loop between AI systems and society that often remains overlooked and under-researched. Algorithms on platforms like TikTok, Amazon, and Airbnb influence mainstream opinions, political support, travel habits, job market outcomes, and product popularity. These systems shape societal behaviors, the evolution of business models, and the emergence of new platforms.

I aim to work with sociologists, political scientists, and psychologists to understand these dynamics better. The final goal is to identify how democratic and digital institutions can guide these influences toward benefiting the common good and mitigating negative impacts.

**Concrete Plans for Research Projects**

I am open to participating in faculty-wide or cross-faculty efforts to acquire third-party funding. However, I have already planned concrete projects that I would like to tackle in the future. Some of these projects depend on the success of this application.

In the following, I present eight project endeavors in different stages of development. All projects with potential collaborators have seen initial discussions and expressions of interest. For each, I identify a potential funding source, relate it to the previously discussed research directions, suggest a time frame, outline the core conception and current progress, and provide an initial funding estimate, including my expected share.

**Operationalizing Effective Human Oversight (E8)**

**(CPEC Subproject, Direction A, time frame: mid 2024)**

*Context and Core Idea:* While research into the conceptual prerequisites for effective human oversight is growing, the practical differentiation between effective (or meaningful) and merely nominal human oversight has largely been neglected. The central focus of E8 involves the operationalization and certification of effective human oversight – a task that presents an inherently interdisciplinary challenge that necessitates bridging technological feasibility with conceptual foundations, normative requirements, and empirical methodologies.

*Status:* The project is requested by the CPEC steering committee for integration into the current phase and is meant to prepare the ground for future projects for the next phase of CPEC (see below). E8 will be developed together with Markus Langer, who is already a PI in CPEC and will take over the empirical side of the research. The details of the E8 proposal are currently finalized with the CPEC steering committee.

*Funding:* Estimated at **250,000 euros**, with two-thirds allocated to me (covering a half position for about 30 months).

**Trusted Health.AI (THAI) – Certification of Trustworthy AI-based Health Solutions**

**(in collaboration with BSI, Direction A, time frame: late 2024)**

*Context and Core Idea:* An AI System used in the healthcare sector is, by definition, high-risk (cf. EU AI Act). This results in a large number of normative requirements (regulatory but also arguably moral ones) that are typically subsumed under the label »Trustworthy AI«, calling for suitable certifications. While proposals for certification exist, the specifics of the process and conditions for success are not well-defined. THAI aims to establish a comprehensive, application-focused certification process for AI-based healthcare solutions.

*Status:* An initial proposal has been submitted to the *Health.AI* network's third call, led by a consortium including DFKI/CERTAIN (represented by me), Prof. Markus Langer (Professor for work and organizational psychology at the University of Freiburg) and the companies QuantPi and ki:elements. After an invitation to pitch, the consortium was invited to submit a full proposal,

but it was revealed that Health.AI lacked sufficient funding for the project. A meeting with the BSI in April will discuss potential alternative project funding through a BSI-specific call.

*Funding:* Estimated at **1 to 1.5 million euros** over 2.5 to 3 years, with significant portions allocated to DFKI and the chair, focusing on practical research and trustworthiness certification. The chair's research will particularly address accountability and informed consent. Approximately **375,000 to 450,000 euros** for DFKI and **180,000 to 250,000 euros** for the chair are aimed at; remaining funds will support other partners.

#### ***Explainable Intelligent Systems 2.0: Next-Generation XAI in Good Practice***

(potentially **EIC Pathfinder Open** or other, **Direction A**, time frame: **mid 2025**)

*Context and Core Idea:* Building on the success of the EIS project, this follow-up aims to delve deeper into model-specific XAI methods, particularly those exploring the mechanistic interpretability of artificial neural networks. It focuses on creating more reliable and robust explanations, hopefully allowing the formulation of better guarantees and helping to ensure that AI systems are ethically sound and practically applicable. The project also addresses the risk of misuse by enhancing our understanding of knowledge extraction.

*Status:* As a natural continuation and expansion of the EIS project's research, this project will further explore the intersection with several other existing and planned efforts. The planned collaboration includes Prof. Lena Kästner and Dr. Timo Speith, experts in cognitive architectures in natural and artificially intelligent systems and interdisciplinary XAI research. The project aligns with activities within CERTAIN and has synergy potential with the RTG Neuro-Explicit Models, the NextAID project and the corresponding Cluster of Excellence proposal NextAID<sup>3</sup>, especially regarding the trustworthiness of AI models in the context of AI-supported drug discovery and addressing dual-use concerns.

*Funding:* Targeting a budget comparable to EIS, with **1 to 1.5 million euros** envisioned. Of this, **300,000 to 400,000 euros** would support the work at a DFKI research department and the chair for the ethics of digitalization.

#### ***Ethics in the Loop (EitL) – Interlocking Normative Requirements and Perspicuity Properties for Responsible Hybrid Decision-Making***

(**SAB Proposal + BMDV, Direction A**, time frame: **early and mid 2025**)

*Context and Core Idea:* With the establishment of a research department at DFKI, a primary objective is to utilize the BMBF »Rahmenvertragsmittel« which entails, first and foremost, gaining the approval of DFKI's *Scientific Advisory Board* (SAB, refer to MAC-MERLin in the third-party funding document). Building on preliminary efforts and projects focused on effective human oversight and responsible/autonomous hybrid decision-making, the EitL project seeks to make these concepts operational by developing design and monitoring frameworks, enhanced by empirical research. EitL explores how contextual factors relate to these desiderata and the technological feasibility with respect to explainability (XAI) and bias detection. Thus, it serves as both a practice-oriented continuation of the EIS project and an expansion of E8 (and, pending application success, THAI), laying further groundwork for the third phase of CPEC (see below).

*Status:* At the recent SAB meeting, I engaged in unofficial discussions with SAB members about the EitL project idea. The concepts received positive feedback, particularly regarding the potential to enhance trust calibration and the assessment of AI system trustworthiness in a measurable manner. This direction also complements the initiatives planned in WP4 of MAC-MERLin (see attached document on third-party funding).

*Funding:* Typically seeking **1.6 to 1.8 million euros** over 3 years for staffing, materials, and travel. Additionally, preliminary discussions with the *Federal Ministry for Digital and Transport Affairs* (BMDV) have shown interest in supporting the research endeavor in the context of the pending integration and implementation of the EU AI Act's Article 14 into national law. BMDV indicated the possibility of contributing **800,000 to 1 million euros** annually, with significant funding allocated to the research department but also to collaborations from the law. Initial preliminary consultations with Dr. Andreas Sesing-Wagenfeil have taken place.

### **On Good Models**

(ERC Starting Grant, Direction B, time frame: mid 2026)

*Context and Core Idea:* My research of the last years has focused on various quality measures, particularly in relation to XAI methods. In the next research phase, I will explore the relationship between quality measures, investigating their hierarchies and potential trade-offs, such as between accuracy and robustness versus fairness and explainability. The project will, for instance, examine the mathematical incompatibilities among fairness definitions and the balance between usefulness and truthfulness in explanations, the influence of contextual factors, like the application field, but also the embedding within larger decision structures, on the hierarchy of such quality measures and on such trade-offs.

*Status:* Preparatory work includes systematic review and presentations on this topic, laying the groundwork for an ERC Starting Grant application in the next two years.

*Funding:* 1.5 million euros for a 5-year period.

### **Philosophically Informed Approaches to Machine Ethics and AI Alignment**

(potentially ERC Synergy Grant or other, Direction C, time frame: early 2026)

*Context and Core Idea:* The intersection of machine ethics and AI has garnered increasing focus, with philosophical and technical perspectives often at odds. This project aims to bridge this gap by developing a comprehensive philosophical background for machine ethics and a formal ethical framework for AI, particularly within reinforcement learning contexts. These efforts seek to harmonize ethical principles with AI decision-making processes.

*Status:* My involvement in Dagstuhl seminars and the establishment of a trans-European network underscore the field's interest in these topics. Collaborations with experts across Europe and the US are planned, including Prof. Michael Fisher, Prof. Louise Dennis (both Manchester, UK), Prof. Marija Slavkovik (Bergen, Norway), Dr. Adriano Mannino (Berkeley, US), Dr. Aleks Knoks (Luxembourg), and Dr. Timo Speith (Bayreuth, Germany). Recent publications and other contributions to academic conferences, including my invitation to program committees (AA-MAS, EUMAS, Formal Ethics, IJCAI, see also CV), underscore the potential of the project idea.

*Funding:* Too early to tell.

### **CRC Spin-offs from Unsuccessful Anthropic Informatics Proposal**

(DFG CRC, Direction B and D, time frame: 2025 to 2027)

*Context and Core Idea:* The Cluster of Excellence undertaking Anthropic Informatics of the Saarland Informatics Campus has inspired new research directions. I am particularly drawn to two emerging ideas: exploring practical approaches to justice beyond conventional fairness operationalizations and understanding and managing societal-scale feedback loops.

*Status:* A promising consortium is already forming around the latter topic, including Prof. Anja Feldmann, Krishna P. Gummadi, Prof. Ingmar Weber, and Prof. Christoph Sorge, with recent talks pointing towards a collaboration with the Interdisciplinary Institute for Societal Computing (I2SC). The potential for a CRC proposal is currently being explored.

Concerning the first direction, there have already been initial discussions (especially with Prof. Isabel Valera), but these seem, for now, to amount to joint preparatory work rather than concrete project plans.

*Funding:* Too early to tell.

### **Ethics of Perspicuity**

(DFG CRC (third phase of CPEC), Directions A, B, C, and D, time frame: 2027)

*Context and Core Idea:* As CPEC progresses to its third phase in 2027, the focus shifts towards interdisciplinary approaches and practical applications. Based on my involvement in the initial phases and a forthcoming PI role (see E8 above and E7 in the attached third-party funding document), this project will explore the ethics of perspicuity at both individual (for instance, effective human oversight, i.e., an application-oriented continuation of E7 and E8, Direction

**A)** and societal levels (**Direction D**). However, the topics may well concern other quality measures of good models and the task of explicating design decisions regarding their hierarchy and trade-offs (**Direction B**). Further, machine ethics may also be addressed, for instance, by making the design decisions of developers and manufacturers accessible to a democratic discourse through explication, which matters especially in light of the lack of agreement on moral background theories (**Direction C**).

*Status:* My future role within CPEC has been discussed within both the E7 consortium and broader CPEC community, reflecting a shared interest in integrating ethical perspectives into perspicuity research, potentially in multiple sub-projects. With E8, further preparatory work is to be carried out in the near future as part of the current CPEC phase. So, while the exact structure has to be the subject of future application writing, it seems foreseeable that I will be involved in the third phase, especially if I remain at the Saarland Informatics Campus.

*Funding:* Too early to tell.

## On Further Ventures

---

I see two further and more general opportunities for cooperation, projects and third-party funding that I would like to mention.

**Teaching Ethics of Digitalization and AI.** The question of how the ethics of digitalization, in general, and AI ethics, in particular, can be effectively taught remains under-explored. In recent years, many such courses have been set up on an ad hoc basis, and the supra-regional visibility that *Ethics for Nerds* has resulted in many inquiries on the course materials, but also the rationale of the course structure. There is a real need for networking and exchange, and the question of how the success of such teaching endeavors, relative to various specific target groups, can be measured and scientifically evaluated. There is definitely a research question as well as funding potential.

**Transferlab for Continuous Ethical Monitoring.** Since I've been at DFKI, I have been repeatedly asked how the ethical quality of AI systems can be assessed in practice. Two factors make it unrealistic to hope for a simple checklist, benchmark, or test bed: Firstly, the fact that moral evaluation often depends on how certain systems develop over time and what effects they have in practice once they are deployed. In this respect, empirical, future-oriented questions stand in the way of an *ex-ante* assessment, a fact that calls for a kind of independent, continuous ethical monitoring. Secondly, hardly any provider of AI systems wants to grant full access to their generally proprietary systems.

I believe these observations offer a promising starting point for designing a DFKI *Transferlab*, i.e., a structured sub-unit for researching and experimentally testing innovative processes in cooperation with partners from application and industry, designed for the medium to long term.

## Structural Prerequisites

---

The prerequisite for the implementation of the project ideas outlined above and for the successful exploration of the envisaged research directions is the successive establishment of a team of highly motivated and talented scientists at various stages of their careers who are passionate about interdisciplinary research. Based on my experience over the past few years, I am convinced that the Uds and the *Saarland Informatics Campus* offer the perfect conditions precisely in this respect. In particular, *Ethics for Nerds* has proven to be an excellent crystallization point, even a hotbed of talent, in recent years. Only in this manner was I able to foster and retain so many young talents. An expansion of such teaching activities and the development of a possible inherently interdisciplinary study program (see the attached «Teaching Statement») could help to further leverage and expand this potential. Saarland could thus become a real beacon for genuine interdisciplinary research in the field of the ethics of digitalization and AI ethics – in Germany, the European Union, and even worldwide.

## Teaching Statement

---

**Teaching Goals.** Beyond imparting basic philosophical knowledge, I prioritize teaching core philosophical methods, such as precise and clear argumentative thinking and writing. These skills empower students across *all* disciplines to effectively articulate and justify their perspectives. Inspiring and motivating my students is at the heart of my approach.

**Teaching Methodology.** I embrace interactive and inclusive teaching practices. To foster engagement, I encourage student-led presentations and utilize digital tools like polls and self-tests. Following the *Good Practice Guide by the Society for Women in Philosophy (SWIP)*, I ensure our classroom is a space where every voice can contribute, employing strategies like small group discussions and structured, text-based preparatory work to diversify participation.

**Grading.** I am committed to transparency and balance. I set clear criteria in a handout at the beginning of each course and try to supervise homework, term papers, and theses so there is enough room for personal development.

**Evaluation and Feedback.** I believe in the value of iterative, evaluation-driven improvements in course design. This approach transcends mere content updates – though those remain essential, especially for the courses I typically give – and delves into refining a course's methodological emphasis and overarching direction. The development of *Ethics for Nerds* illustrates this process well. Initially, for example, I assumed computer science students' proficiency in reasoning and logic would align with the course's focus on moral philosophical tenets over argumentation's intricacies. However, ongoing evaluation, continuous feedback, and reflective dialogue highlighted a crucial incapability of computer science students to reason formally within the context of natural language. This insight led us to significantly bolster our coverage of *critical thinking*. This adjustment has not only led to introducing a critical thinking segment into the course but transformed it into a highlight, celebrated for its impact and value by our students.

## Concrete Teaching Ventures

---

**Philosophy.** I am eager to diversify further my teaching portfolio, which includes integrating aspects of *political philosophy* and *philosophy of law* with digital and AI-related relevance. Further, I aim to explore topics such as *informed consent*, the interplay between *epistemic norms and ethical requirements*, the impact of *digitization on sustainability*, and the ongoing debate around *AI alignment*. I am also well-prepared to contribute to foundational courses, including *Introduction to Practical Philosophy* and *Introduction to Ethics*, if desired.

**Ethics for Nerds.** Should I have the privilege of being in charge of *Ethics for Nerds* again, I would prioritize its continuation and advancement. I consider splitting it into a *core lecture* and an *advanced lecture* on AI-Ethics and envisage creating specialized lectures that integrate AI, philosophy, and ethics, such as an *>Ethics for Nerds< for Philosophy Geeks* course designed for philosophy students. Moreover, there is a notable demand for variation of *Ethics for Nerds* for various postgraduate and certificate programs, including the Master of Laws (LL.M.) in *Information Technology and Law, Digitalität. Kl. Gesellschaft*, and the *Transform4Europe* focus in the *Europa-icum*. In this spirit, I would also love to develop versions of *Ethics for Nerds* specifically tailored to different audiences within the *Saarland Informatics Campus*, such as senior researchers.

**Informatics and DFKI.** I am open to creating new philosophically-informed *seminars* focusing on fairness and explainability, particularly tailored to CS students. In partnership with the DFKI, I am keen to offer seminars on *responsible AI, regulation ethics, and (formal) machine ethics*. Furthermore, I am excited about the prospect of starting a regular interdisciplinary seminar on AI for the common good. I also look forward to contributing to the summer school initiated by Prof. Sebastian Vollmer at DFKI Kaiserslautern.

**Informatics and DFKI.** Finally, my experience has underscored the essential need for individuals with profound interdisciplinary expertise in philosophy and informatics. With this in mind, I would be enthusiastic about an opportunity to design and, potentially, lead a study program that seamlessly integrates these critical disciplines.



# Doing Wrong with Others

MULTI-AGENT CONSEQUENTIALISM

AS A SOLUTION FOR THE COLLECTIVE ACTION PROBLEM

by

Kevin Baum

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Submitted to the Faculty of Humanities and Theology

Department of Philosophy and Political Science

FIRST REVIEWER: Prof. Dr. Eva Schmidt (TU Dortmund)

SECOND REVIEWER: Assoc. Prof. Dr. Vuko Andrić (Linköping University)

## Abstract

According to MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (MOAC), more a family of theories than a specific doctrine, the concepts of the *right* and the *best* are closely intertwined. MOAC theories assert that an action is right if and only if no alternative action has better consequences. This *criterion of rightness* seems, however, to be an expression of a more general view, according to which the ›core function of morality‹ consists in implicitly coordinating collective actions: Those actions that, if carried out, lead to the morally best world that moral agents can collectively bring about are to be designated as right. This idea, prominently referred to as PRINCIPLE OF MORAL HARMONY by Fred Feldman (Feldman [1980]), was considered unchallenged dogma within the consequentialist community until the second half of the 20th century (for example, see Bentham [1780]; Baier [1958]; Castañeda [1974]; Mackie [1977]).

However, whether MOAC theories can meet this expectation is questionable. For various structures – overdetermination and preemption, as well as the apparent existence of effects that, considered in isolation, are negligible but accumulate into significant harm – seem to allow the existence of collective decision situations in which combinations of actions yield collectively suboptimal results, although no agent could have made a difference for the better by acting differently unilaterally. Consequently, such actions are apparently right according to MOAC, *although* they lead to suboptimal outcomes. This puzzle, known as the CHALLENGE OF COLLECTIVE ACTION, has challenged consequentialism for decades (to name just a few, see Glover and Scott-Taggart [1975]; Regan [1980]; Parfit [1984]; Zimmerman [1996]; Kagan [2011]; Pinkert [2015]).

This dissertation aims to reconstruct and understand the CHALLENGE OF COLLECTIVE ACTION in its various forms and ultimately propose a novel, consequentialist solution. This solution, significantly based on game-theoretical considerations, results in a generalized, maximizing objective consequentialist theory for multiple agents named MULTI-AGENT CONSEQUENTIALISM. The overarching aim of this work is to preserve PRINCIPLE OF MORAL HARMONY as a fundamental motivation of consequentialist theorizing and, at the same time, to offer a decidedly objective-consequentialist solution to the CHALLENGE OF COLLECTIVE ACTION.

## Abstract

Gemäß MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (MOAC), eher eine Familie von Theorien als eine spezifische Doktrin, sind die Begriffe des *Richtigen* und des *Besten* eng miteinander verwoben. MOAC-Theorien behaupten, dass eine Handlung genau dann richtig ist, wenn keine alternative Handlung bessere Konsequenzen hat. Dieses *Kriterium der Richtigkeit* scheint indes Ausdruck einer allgemeineren Ansicht, wonach die ›Kernfunktion der Moral‹ darin besteht, kollektive Handlungen implizit zu koordinieren: Es gilt diejenigen Handlungsoptionen als richtig auszuweisen, die, wenn ausgeführt, zur moralisch bestmöglichen Welt führen, die die moralischen Akteure zusammen herbeizuführen vermögen. Diese Idee, die von Fred Feldman primärweise als PRINCIPLE OF MORAL HARMONY (Feldman 1980) bezeichnet wurde, galt bis zur zweiten Hälfte des 20. Jahrhunderts innerhalb der konsequentialistischen Gemeinschaft als unangefochtenes Dogma (beispielsweise siehe Bentham 1780; Baier 1958; Castañeda 1974; Mackie 1977).

Doch ob MOAC-Theorien dieser Erwartung gerecht werden können, ist fraglich. Denn verschiedenste Strukturen – Überdetermination und Präemption sowie die scheinbare Existenz von Effekten, die isoliert betrachtet vernachlässigbar erscheinen, sich jedoch zu großen Schäden akkumulieren können – scheinen die Existenz kollektiver Entscheidungssituationen zu erlauben, in denen Kombinationen von Handlungen kollektiv suboptimale Ergebnisse liefern, obgleich kein Akteur durch einseitiges Andershandeln einen Unterschied zum Besseren hätte machen können. Folglich scheinen solche Handlungen gemäß MOAC richtig zu sein, *obwohl* sie zu suboptimalen Ergebnissen führen. Dieses Rätsel, bekannt als das CHALLENGE OF COLLECTIVE ACTION, hat den Konsequentialismus seit Jahrzehnten herausgefordert (um nur einige zu nennen, siehe Glover and Scott-Taggart 1975; Regan 1980; Parfit 1984; Zimmerman 1996; Kagan 2011; Pinkert 2015).

Ziel dieser Dissertation ist es, das CHALLENGE OF COLLECTIVE ACTION in seinen verschiedenen Formen zu rekonstruieren und zu verstehen, um schließlich eine neuartige, konsequentialistische Lösung vorzuschlagen. Jene signifikant auf spieltheoretischen Überlegungen basierende Lösung führt zu einer verallgemeinerten, auf Maximierung ausgerichteten konsequentialistischen Theorie für multiple Akteure namens MULTI-AGENT CONSEQUENTIALISM. Das übergeordnete Ziel dieser Arbeit ist es, PRINCIPLE OF MORAL HARMONY als grundlegende Motivation der konsequentialistischen Theoriebildung zu bewahren und gleichzeitig eine dezidiert objektiv-konsequentialistische Lösung für das CHALLENGE OF COLLECTIVE ACTION anzubieten.



# Preface

In February 2022, when President Putin initiated an illegitimate war against Ukraine, Germany found itself at a crucial juncture. Decades of increasing reliance on Russian gas had left the country vulnerable to potential supply crises. Just days after the invasion, a fervent debate erupted regarding potential shortages of this vital resource.

As the situation intensified, appeals to citizens to take individual responsibility became ubiquitous. On social media, posts like this tweet<sup>1</sup> by Rico Grimm, a German journalist, were popular: »What can you do against Putin? It's never been easier: turn down the heating, put on a sweater.« Echoing this sentiment, Robert Habeck, Germany's Federal Minister for Economic Affairs and Climate Action, sounded the alarm on March 30. Triggering the early warning stage for gas supplies, he reached out to the German populace with<sup>2</sup> a resonant plea:

We are in a situation where I have to say clearly that every kilowatt hour of energy saved helps, and that is why I would like to combine the triggering of the early warning level for gas supplies with an appeal for help to companies and private consumers: You are helping Germany, you are helping Ukraine when you reduce your use of gas, or energy in general.

---

<sup>1</sup>[https://twitter.com/gri\\_mm/status/1499368527844757516](https://twitter.com/gri_mm/status/1499368527844757516), my translation.

<sup>2</sup>Cf. <https://www.dw.com/en/german-economy-minister-raises-warning-level-for-gas-supplies/a-61300264>

Robert Habeck's call was not to Germany as an entity but to its individual citizens. He urged them to evaluate and potentially adjust their behaviors in light of the looming gas shortage and its associated repercussions. The underlying message was clear: every little individual contribution, no matter how small, helps.

However, skepticism quickly followed. Critics questioned the feasibility and sensibility of a ›freezing for peace‹ approach. They labeled such an appeal as potentially arrogant and cynical.<sup>3</sup> They pointed to the vast capacity of gas storage, the adaptability of markets, and the minuscule proportion of private consumption when juxtaposed against industrial demands. The common argument was this: Individual efforts were like a drop in the ocean. So, while most agreed that it was indeed crucial for Germany and the broader EU to cut down on consumption and move away from Russian gas dependency, they viewed individual efforts as barely consequential.<sup>4</sup> Thus, how can individuals be expected to shoulder the burden of past political misjudgments – especially given that the behavior of the individuals seems so negligible?

This thesis is motivated by the question of the validity of the skeptics' stance. If they're correct, it means that although *we* – considered here as a rather loose collective of rather uncoordinated individual citizens – are morally required (in some sense) to take *collective action* (like reducing gas consumption), each *individual* might *justifiably* abstain because their isolated

---

<sup>3</sup>Cf. [https://www.t-online.de/finanzen/unternehmen-verbraucher/id\\_91797770/frieren-fuer-frieden-das-zeugt-von-purer-arroganz.html](https://www.t-online.de/finanzen/unternehmen-verbraucher/id_91797770/frieren-fuer-frieden-das-zeugt-von-purer-arroganz.html)

<sup>4</sup>This argument is replicated in the March first episode of the podcast *Deutschlandfunk: Der Tag* (<https://www.deutschlandfunk.de/01-03-2022-der-tag-blau-gelb-im-europaparlament-d1f-21205766-100.html>) and the *Die Zeit: Energiesparen gegen Putin* article on March 13 (<https://www.zeit.de/wirtschaft/2022-03/gasversorgung-russland-ukraine-krieg-embargo-energiesparen>).

efforts may not lead to any significant contribution. Thus, the question driving this project is:

- (Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could bring about together?

For the past decade, this question has occupied my mind. I think I was particularly attracted to (Q) because, although it is a *theoretical question*, it underlies many *practical, real-life challenges*, some of which are among the greatest of our time. For instance, the man-made climate crisis can be discussed morally on several levels against the background of (Q) (cf. Sinnott-Armstrong and Howarth [2005], Gardiner [2011], Budolfson, McPherson, and Plunkett [2021]), as can questions of moral obligation to resist oppressive social systems or questions of vaccination. Thus, while this inquiry primarily probes a theoretical, moral-philosophical dilemma, its implications are far from purely theoretical. In fact, it is of as much practical importance as moral questions can be.

This thesis addresses (Q) as a theoretical challenge, though. It does so from the perspective of a specific notorious family of moral theories for which that challenge is arguably particularly pressing, namely MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (or simply MOAC for short). By the end of my investigation, I will have shown that, contrary to popular belief, an advanced MOAC theory can affirm (Q) for a broad class of relevant cases. However, I will also argue that (Q) *must* be denied for another class of critical but particularly relevant cases but that this is not a deal-breaker for MOAC.

## On Style and Methodology

I am convinced that scientific progress is possible in the humanities, including philosophy. But it comes with the necessity to wrestle with thoughts, words, and concepts not only until what is to be said is said – and no more and no less – but also until what was complicated to understand is easy to understand. In this regard, I subscribe to the ideals of analytical philosophy. Precision in thought and language, structure and systematicity, arguments and conceptual analysis, and a readiness to apply formalism where useful or even necessary are essential features of doing good philosophy.

Meeting these demands is not easy. The path to initial knowledge is winding and is never the shortest and straightest path. A better and more convenient path, in general, can be found, however, once you have reached your destination – provided you make the effort. The task of the scientific philosopher is to show others such a better and more convenient path, or at least a suitable approximation, possibly taking into account the one or the other lookout or sight worth seeing that one found on one's own way to the top. It is certainly not good philosophy to force the readers to take the original or another unnecessarily complicated and stressful path.

While I have been struggling to find such a convenient path, I think I found an acceptable one. As is so often the case, however, one is never fully satisfied at the end of such an endeavor. For sure, you can always find another better path. Where I have failed, I hope to have failed, at least in an engaging and instructive way.

The purpose of this book is to help MOAC, one of the most important and influential families of moral theories, by addressing one of its most pressing

internal theoretical challenges. To do so requires drawing on the (often implicitly agreed on) conceptual toolbox of this family of theories and expanding it here and there. Assertions and argumentative attacks must be understood well to be able to reject them. For this, formalism can be a necessary tool. Part of my challenge, thus, was to ›go formal‹ when it adds precision, clarity, and insight and eschew formalism when it threatens to obscure understanding. Achieving this balance has been a delicate act, and while I've occasionally faltered, I hope, more often than not, I've succeeded.

However, this could have easily become another project. Many of this book's core ideas and claims could certainly be proved and derived within a *thorough* formal framework. But that would then show, first of all, that the ideas I hold and the points I make are valid *within* that specific formal system. Until it is shown that the relevant system captures what we (or at least the champions of camp MOAC) call morality, however, this would not show what I want to show. Such a project would be one of deontic- or, more precisely, multi-modal- logic. At times, my project was on the verge of becoming this project, which would undoubtedly be an exciting project in its own right. But it is neither the project I started with nor the project I ended with. It is not this project.

In this work, as a consequence, I employ what can best be described as a kind of semi-formalism. The used notations, notions, concepts, and structure, while resonating with known frameworks, are distinct from existing ones but well-suited for my purposes. If I were to align it more closely with any established framework, stit semantics would be the chosen one (where »stit« stands for »seeing to it that«, cf. Belnap, Perloff, and Xu [2001]; Harty and Belnap [1995]; Harty [2001]). Stit semantics offers a structured platform for

multi-modal logics of action, drawing inspiration from Arthur Prior's branching time model (Arthur N. Prior [1967]; Arthur N. Prior [1955]). In many ways, the formalism presented in this thesis can be perceived as stit ›undercover‹, enabling the utilization of stit's core concepts without being constrained by its rigorous formalities. I owe a profound debt of gratitude to John Horty's seminal works, especially his in-depth explorations in the field (cf. Horty [2001]; Horty [2019]).

## Acknowledgements

Many people have supported my academic path in general and this project in particular, in so many different ways! I am deeply grateful to all of them and will articulate that gratitude more emphatically and in more detail on another occasion.

# Contents

<b>Preface</b>	v
<b>1 Introduction</b>	1
1.1 The CHALLENGE in a Nutshell . . . . .	6
1.2 Structure . . . . .	15
<b>I The CHALLENGE and How Not to Solve It</b>	17
<b>2 Preliminaries I</b>	21
2.1 Individual Decision Situations . . . . .	21
2.2 Moral Theories, the Rightness Predicate . . . . .	27
2.3 MOAC and Its Three Modules . . . . .	37
2.3.1 Relevance Stances . . . . .	40
2.3.2 Criteria of Rightness . . . . .	44
2.3.3 Axiological Background Theories . . . . .	46
2.3.4 Putting It All Together . . . . .	51
<b>3 The CHALLENGE</b>	57
3.1 On Choosing Giants . . . . .	57
3.2 The PYRAMID . . . . .	60

<b>3.3 The TRILEMMA . . . . .</b>	<b>62</b>
<b>3.4 The CHALLENGE as No-DIFFERENCE CHALLENGE . . . . .</b>	<b>71</b>
<b>3.5 The CHALLENGE as CHALLENGE<sub>int</sub> . . . . .</b>	<b>98</b>
<b>    3.5.1 Starting from a Non-Starter:</b>	
The Principle of False Universalization . . . . .	98
<b>    3.5.2 Regan's Impossibility Result . . . . .</b>	<b>106</b>
3.5.2.1 Step 1: WHIFF AND POOF and $P_{\exists T}$ . . . . .	111
3.5.2.2 Step 2: PROPCOP and $P_{MH}$ . . . . .	113
3.5.2.3 Step 3: Regan's ›Proof‹ and $P_{MOCOr}$ . . . . .	123
<b>    3.6 The PYRAMID and the Next Steps . . . . .</b>	<b>145</b>
<b>4 Good Solutions, Bad Solutions, Non-Solutions</b>	<b>147</b>
<b>    4.1 Mapping the Solution Space . . . . .</b>	<b>148</b>
<b>    4.2 A Limitation: On the Exclusion of CUMULATIVE EFFECTS</b>	
CASES From the Scope of this Project . . . . .	156
<b>    4.3 Criteria for Good Solutions . . . . .</b>	<b>162</b>
4.3.1 Requirements: . . . . .	164
4.3.2 Cachets: . . . . .	166
4.3.3 Some Notes on Criteria . . . . .	172
4.3.4 A Counterexample Against DEONTIC COMPLETE-	
NESS? . . . . .	173
<b>    4.4 An Unsatisfying Exploration of the Solution Space . . . . .</b>	<b>178</b>
4.4.1 Kagan's Revived Discourse . . . . .	180
4.4.2 Pinkert: Modal Virtue Consequentialism . . . . .	183
4.4.3 Jackson: Collectivism . . . . .	186

CONTENTS	xiii
----------	------

<b>II The REAL CHALLENGE and How to Solve It</b>	<b>193</b>
<b>5 Preliminaries II</b>	<b>199</b>
5.1 Collective Decision Situations . . . . .	200
5.2 (Semi) Formalism and Shorthands . . . . .	207
5.2.1 Domains and Properties and the Triad . . . . .	210
5.2.1.1 Maximality . . . . .	212
5.2.1.2 Order-Invariance . . . . .	213
5.2.1.3 Symmetry . . . . .	217
5.2.2 1-Variants and Independence . . . . .	223
5.3 Revisiting SEQUENTIAL CASES . . . . .	224
5.4 Separability and Conditionalization . . . . .	226
5.4.1 Formal Toolbox for Reductions . . . . .	234
<b>6 The REAL CHALLENGE</b>	<b>239</b>
6.1 PRINCIPLE OF MORAL BALANCE . . . . .	240
6.2 Revisiting the ARGUMENT . . . . .	245
6.3 The Intuition: Gaps Filled Badly . . . . .	246
6.4 The Logical Structure of The ARGUMENT . . . . .	255
6.4.1 The Structure of $P_{\exists T}$ : Straightforward . . . . .	256
6.4.2 The Structure of $P_{MOCOR}$ : Ex Post! . . . . .	260
6.4.3 The Structure of $P_{MH}$ : Ex Ante . . . . .	261
6.4.4 Putting Things Together . . . . .	263
6.5 When The Villain Finally Enters The Stage:	
The REAL CHALLENGE . . . . .	265
<b>7 Of New Consequences</b>	<b>277</b>

<b>7.1 New Grounds for Consequentialism . . . . .</b>	<b>278</b>
<b>7.1.1 »Like Scales Fell From His Eyes...« . . . . .</b>	<b>279</b>
<b>7.1.2 Exotic and Esoteric? Or Old Wine in New Bottles? . . . . .</b>	<b>285</b>
<b>7.2 Towards a Unified Representation: . . . . .</b>	
<b>The Generalized Extensive Form . . . . .</b>	<b>291</b>
<b>7.3 Filling Gaps With Multi-Agent Amendments . . . . .</b>	<b>301</b>
<b>7.3.1 Aggregative Approaches . . . . .</b>	<b>303</b>
<b>7.3.1.1 SUMMATION . . . . .</b>	<b>305</b>
<b>7.3.1.2 MAXIMIZATION . . . . .</b>	<b>309</b>
<b>7.3.1.3 EXPECTED UTILITY . . . . .</b>	<b>312</b>
<b>7.3.2 Non-Aggregative Amendments . . . . .</b>	<b>316</b>
<b>7.3.2.1 (Non-)Domination . . . . .</b>	<b>317</b>
<b>7.3.2.2 MAXMIN and MAXMAX . . . . .</b>	<b>319</b>
<b>7.3.2.3 MIXED STRATEGIES . . . . .</b>	<b>321</b>
<b>7.4 What Remains to Be Done . . . . .</b>	<b>327</b>
<b>8 On Reasonable Disharmonies and The Quest for . . . . .</b>	
<b>    The Best Amendment . . . . .</b>	<b>329</b>
<b>8.1 Revisiting PMH . . . . .</b>	<b>330</b>
<b>8.1.1 The Limits of PMH . . . . .</b>	<b>332</b>
<b>8.1.2 Upshot: Reasonable MORAL HARMONY . . . . .</b>	<b>342</b>
<b>8.2 Amendments for Reasonable Pathfinding . . . . .</b>	<b>343</b>
<b>8.2.1 On Policies and Their Evaluation . . . . .</b>	<b>344</b>
<b>8.2.2 Evaluating Amendments . . . . .</b>	<b>348</b>
<b>8.3 The Final Evaluation . . . . .</b>	<b>352</b>
<b>8.3.1 Defining a Testbed . . . . .</b>	<b>354</b>

CONTENTS	xv
8.3.2 And the Winner Is ... . . . . .	366
8.4 On Overall Success . . . . .	370
<b>9 Summary and Future Work</b>	<b>379</b>
9.1 Future Work? . . . . .	382
9.1.1 Implications for Subjective Consequentialism . . . . .	382
9.1.2 Implications for The Actualism/Possibilism Debate .	383
9.1.3 Generalizaion . . . . .	387
9.1.4 Formal Proofs . . . . .	390
9.2 How the Tables Have Turned . . . . .	391
<b>Bibliography</b>	<b>397</b>



# Chapter 1

## Introduction

The relationship between the *morality of actions* and the *moral quality of their consequences* is a topic of much debate. A significant portion of the philosophical community even believes that the *primary function* of morality is to guide us towards morally optimal outcomes. Thus, they typically find themselves aligned with the following principle:

- (R) If someone does what is morally right, they are guaranteed to produce the morally best possible results that they could bring about.

The topic at the core of the present project is whether this intricate relationship translates to *collective contexts*. This boils down to the following question:

- (Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could bring about together?

Those who believe in (R) will arguably be inclined to affirm (Q). Of course, other conceptualizations of morality do *not* emphasize the relationship between the right and the best so much. Thus, those proponents will often readily deny (Q) without second thought. This research, however, is tailored for those who see value in the first viewpoint.

It has been argued for decades that, for a particular, highly influential family of moral theories, as famous as they are notorious, which I will hereafter call **MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM** (or shorter MOAC), denying (Q) is no viable option. *Consequentialist*<sup>5</sup> theories of morality make the moral quality of the consequences of actions (and their alternatives) the *sole* measure of the morality of those actions. *Objective consequentialist* theories limit their theoretical considerations to the moral qualities of the *actual* consequences of actions, i.e., to those consequences that would be the case if a specific alternative was performed. Which consequences are to be expected or are actually expected by some agent is thus irrelevant to objective consequentialist theories. *Maximizing* consequentialist theories ask us to *maximize* the moral qualities and not, for instance, to produce good or sufficiently good consequences (in the sense of suboptimal consequences being good enough, cf. Slote and Pettit [1984]).

As the name implies, MOAC theories are that subclass of consequentialist theories that are objective and maximizing. They are still a *family of theories* – rather than a single, specific theory – in that there is disagreement among MOAC theories about the nature and specifics of the moral quality

---

<sup>5</sup>In the following, I will mostly drop the prefix »act-«. There are other kinds of consequentialist theories, for example, those that operate on the level of motives (Sverdlik [2011] or rules (Brandt [1959], Hooker [2023]). Nevertheless, these do not play a role in this project, so whenever I write of »consequentialist theories«, I always mean *act*-consequentialist theories.

of consequences. It follows that these theories are characterized by a common criterion of *rightness*<sup>6</sup>, the MAXIMIZING OBJECTIVE CRITERION OF RIGHTNESS (or, more briefly, simply MOCoR), which we can tentatively express as follows and that obviously corresponds to (R) above:

**Criterion 1.1 (MOCoR (tentative))** *An action is right if and only if there is no alternative action that would actually lead to better consequences.*

This brings us back to the initial question (Q): On the one hand, MOAC theories seem committed to answering the question *in the affirmative* because they put a consonance between right action and morally optimal outcomes at the center of their philosophical theorizing. On the other hand, specific collective decision scenarios present a conundrum in this regard. This is because, in some situations, the consequences of our actions are inextricably tied to the decisions of others, and this interdependence can often be mutual for all involved. In such cases, all agents can act such that they together bring about suboptimal, even disastrous consequences, while doing so in a way that it seems to be true for everyone that they could not have changed anything for the better by acting differently. In such case, however, MOCoR seems to imply that everyone acts rightly, implying that MOAC theories have to answer (Q) *in the negative*. I call this and similar challenges that we will encounter later the CHALLENGE OF COLLECTIVE ACTION (or simply

---

<sup>6</sup>This work is limited to questions of moral rightness and operates mainly on the widely shared assumption that morally right actions are precisely those actions that are morally permissible (although more needs to and will be said about the precise relationship of these different kinds of moral status). Further, from this point on, I mean by »rightness« by default *moral* rightness. The focus on rightness also largely corresponds to language in a large part of the literature on the CHALLENGE relevant to this project. In some quotations, however, we will encounter talk of an agent who ought to act in specific ways. In these cases, this can be understood in terms of rightness by adopting that an agent ought to perform one of their right alternatives and, thus, that if there is only one right (respectively permissible) option, the agent ought to perform that very alternative.

the CHALLENGE for short). Solving it within the framework of MOAC theories is the primary goal of this work.

The CHALLENGE has occupied moral philosophers for quite some time. At least since the 1970s (cf. Glover and Scott-Taggart [1975]) and in particular in the aftermath of Donald Regan's book *Utilitarianism and Co-operation* (cf. Regan [1980]), consequentialists have endeavored to save MOAC in the face of the CHALLENGE. Some have taken it as an occasion to jettison fundamental consequentialist beliefs (Feldman [1980]; Parfit [1988]; Sinnott-Armstrong [2005]; Jackson [1987]) regarding the (guaranteed) consonance between right action and optimal results. Others used it as a reason for substantial modifications of the consequentialist criterion of rightness (Parfit [1984]; Zimmerman [1996]) or even for abandoning (act-)consequentialist grounds altogether (Regan [1980]). Over the years, the debate has spread to subjective consequentialist terrain (Kagan [2011]; Pinkert [2015]; Portmore [2018]; Budolfson [2019]; Hedden [2020]). Some authors have discussed the CHALLENGE from other, more general perspectives (Andreou [2014]; Nefsky [2011]) and asked how far it extends into non-consequentialist territory (Killoren and Bekka Williams [2013]). This is just a sample of the vast and rich literature on the CHALLENGE.

The presentation so far may give the impression that the CHALLENGE is mainly theoretical. Nevertheless, the CHALLENGE arguably lies at the heart of some of the most pressing practical issues of our time. For instance, think of the anthropogenic climate crisis. Although it is not certain what their concrete form will be, the consequences of our collective greenhouse emissions will undeniably be morally catastrophic. However, it is by no means obvious what this means for the moral status of all the myriad of actions of the individual inhabitants of the earth, which, in sum, are causative for at

least a significant part of those emissions. After all, the emissions of a single, easily avoidable short-distance car trip (or even all the consequences of one individual's consumption and mobility decisions over time) are negligible >globally<. As Walter Sinnott-Armstrong once put it: »global warming and climate change occur on such a massive scale that my individual driving makes no difference to the welfare of anyone« (cf. Sinnott-Armstrong and Howarth 2005).

At the same time, however, avoiding *prima facie* contributing actions is often accompanied by *morally significant* individual costs. Thus, given the apparent individual inefficacy, it seems that while no individual effort to save greenhouse gases will make a morally relevant *positive* difference, many such decisions will make morally significant *negative* ones. This finding is entirely independent of whether, at the end of the day, sufficient greenhouse gas emissions are avoided overall. Therefore, the question arises of how a moral demand for personal sacrifice can then be justified. If a sufficient number of other agents >do their share<, my individual sacrifice is a waste of moral value; and if the collective effort falls short, my individual saving does not change a thing for the better. Thus, it seems morally right (at least in light of MO-CoR) to *not* restrict myself, no matter what the others do. It thus seems better overall if I took that easily avoidable but comfortable car trip. In other words, individual losses appear to trump collective considerations. But then there are situations in which acting rightly leads not only to suboptimal but even to catastrophic consequences. Thus, we are back to (Q) and the CHALLENGE. Therefore, although this project deals with the CHALLENGE in a theoretical setting and in rather abstract terms, it addresses some of the most pressing issues of our time, at least from the point of view of >camp MOAC<.

This project attempts to find a solution to the CHALLENGE that remains true to consequentialist core tenets. Thus, this thesis is a consequentialist project, i.e., limited to a consequentialist perspective. At the same time, the contribution of this work to the advancement of consequentialist theorizing is by no means merely to dispel the CHALLENGE, which is a particular, albeit perilous, and significant challenge for MOAC. For, as will become apparent in the course of this thesis, the CHALLENGE is only a symptom of a more profound failure of objective act-consequentialist theories. Principled and fundamental difficulties arise for instances of MOAC from the fact that the consequences of actions may well depend on the actions of other agents, and the CHALLENGE arises only from the inadequate methods that consequentialists chose for their approach to the CHALLENGE in the past. Therefore, in addition to the specific goal of mastering the CHALLENGE, this work has a more general and theoretically more fundamental goal: it is about nothing less than the search for an objective consequentialist moral theory that can do justice to the fact that each of us is only one among many. Accordingly, this book makes it its mission to develop a consequentialist multi-agent moral theory.

## 1.1 The CHALLENGE in a Nutshell

At its core, this project is an attempt to defend MOAC, and thus a very particular family of moral theories, against a specific challenge: the CHALLENGE. However, there is no canonical representation of the CHALLENGE, but rather several different formulations. These differ, on the one hand, in terms of their potential impact and, on the other hand, in terms of the the-

oretical conditions necessary for their formulation. In the best tradition of analytic philosophy, this work tries to defend MOAC *against the strongest version* of the CHALLENGE for MOAC. However, as it turns out, the strongest version of the CHALLENGE is *not one* variant of it. Instead, it consists of several hierarchically ordered variants that back each other up. The first line of attack consists of a version which I call CHALLENGE<sub>int</sub> (where »int« stands for »internal« which is itself short for »theory-internal«) that operates with a fair amount of, (rather convincing) presuppositions. Its success would invalidate MOAC, establishing the inadequacy of all MOAC theories. If it fails, it can be replaced by other variants, most notably the so-called No-DIFFERENCE CHALLENGE, but also with a pre-theoretic version as a second fallback. These variants get by with significantly fewer strong presuppositions but would also have less dire impacts if successful. Accordingly, to strike back CHALLENGE<sub>int</sub> is the central goal of this thesis. In the following, this variant will be presented in rough sketches, whereby the basic intuitions invoked in the previous section shall get some additional theoretical underpinning. The other two versions will have to wait until Chapter [3], a deep dive into the state of the debate that follows a preliminary chapter (Chapter [2]).

The CHALLENGE<sub>int</sub> is based on the insight that, arguably, MOCoR expresses (or is at least motivated by) a more general view on morality, viz.:

**View 1.1 (CONGRUENCE)** *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

It seems not to be far-fetched that CONGRUENCE is the fundamental assumption underlying MOCoR. In this respect, MOCoR does not just fall

out from thin air but is rather to be understood as a concretization *inspired* by (or, maybe, one possible explication of) CONGRUENCE in the form of a criterion of rightness. MOCOR would, in this case, be intended to capture the spirit of CONGRUENCE and, thus, should be respected to live up to it.

CONGRUENCE and MOCOR not only go well together, but are closely related. That performing one of the actions with the best consequences comes with the best possible consequences seems to be a truism. However, it is also tempting to derive from CONGRUENCE not only a specific criterion of rightness but also read it in a *collective sense*, as an expectation regarding morality to have the general property of ›marking the path to moral optimality‹. In other words, according to such a reading, morality can be expected to highlight exactly those sequences of arbitrary actions of arbitrary agents that, if performed, necessarily lead to (one of) the best possible outcomes that these agents can bring about. Several authors have proposed formulations of this broader claim, which is rather a second-order claim concerning the nature of morality itself: Donald Regan (Regan [1980]), Fred Feldman (Feldman [1980]), who tracked the idea down through history, and, more recently, Felix Pinkert (Pinkert [2015]) and Douglas Portmore (Portmore [2018]). In a first approximation and based on their work, we might capture the idea like this:

### **Principle 1.1 (COLLECTIVELY MAXIMIZING (tentative))**

*If all agents act rightly, then they are guaranteed to produce the morally best outcome they could bring about together.*

MOCOR and COLLECTIVELY MAXIMIZING certainly seem like a perfect match at first sight. It could even be tempting to think that a rightness predicate explicated by MOCOR virtually guarantees the truth of COLLEC-

TIVELY MAXIMIZING. Indeed, one could conclude that if all the agents of some collective perform one of the actions with the morally best possible consequences, this *must* lead to the morally best possible consequences that this collective can bring about together. In coming to such an (as it turns out) *hasty* conclusion, one would be in the best company. Jeremy Bentham, for instance, one of the founding fathers of Utilitarianism and arguably one of the most famous exponents of MOAC (cf. Christopher Woodard [2019]), writes confidently that his *Principle of Utility*, »which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question« (Bentham [1780]), is »capable of being consistently pursued; and it is but tautology to say that the more consistently it is pursued, the better it must ever be for humankind« (cf. *ibid.*).

Based on this seemingly obvious insight, various authors, including several consequentialists, have elevated COLLECTIVELY MAXIMIZING to a touchstone or criterion of adequacy for moral theories. Following Fred Feldman's extensive and thorough work on this ›collective reading‹ of CONGRUENCE (cf. Feldman [1980]), we might call that expectation the PRINCIPLE OF MORAL HARMONY (or shorter just PMH). This name has prevailed in literature to this day (cf. Portmore [2018]). This principle is meant to express the idea that morality ›lights the way‹ to moral optimality not only on an individual but also on a collective level, and that thus, we can make COLLECTIVELY MAXIMIZING a requirement for any potentially adequate moral theory. We can capture<sup>7</sup> the PMH tentatively in terms of a simple necessary condition:

---

<sup>7</sup>I'll use »PMH« to refer to the general idea and use »MH« to refer to different formulations of PMH.

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

Logically equivalent,<sup>8</sup> but slightly reformulated (especially concerning the contraposition applied to the consequence), the statement can also be expressed negatively:

**Criterion 1.3 (MORAL HARMONY (MH, tentative, contraposition))**  
*If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome (i.e., according to this theory), (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).*

Obviously, this takes us back to the question at the beginning of this introduction, which was:

- (Q) If each of us always does what is morally right, are we guaranteed to bring about the morally best results that we could bring about together?

Obviously, if MORAL HARMONY truly is correct, then every adequate moral theory will necessarily affirm (Q).

If Bentham were right, CONGRUENCE, MOCOr, and COLLECTIVELY MAXIMIZING (or, respectively, MH) would fit together perfectly. In particular, COLLECTIVELY MAXIMIZING would simply be an *implication* of MOCOr, which means that the adequacy criterion would trivially be satisfied by

---

<sup>8</sup>The only reformulation here that is not purely syntactic but semantic/conceptual is that I suggest »at least one of the agents acted wrongly« where, strictly speaking, one only needs the weaker »at least one of the agents acted not rightly« is logically warranted. The relevance of this difference will be explored in more detail later.

MOCOR. However, as tempting as it is to elevate an apparent consequence of one's criterion of rightness to a criterion of adequacy for moral theories, it is dangerous when that very adequacy criterion backfires. After all, as so often is the case in philosophy, seemingly obvious, trivial, and uncontroversial statements mark great challenges.

In this case, it is a simple observation that threatens the foundation of consequentialist theorizing: each of us is just one among many, and, moreover, none of us lives in a neatly isolated bubble.<sup>9</sup> We all interact and stand in many kinds of interdependent relationships. Specifically, the results of what we do are usually determined, at least in part, by what others do. Hence, what one person can achieve (or screw up) usually also depends on what others do. However, if the results of our actions depend, at least sometimes and partly, on what others do or will do, then it does not seem settled at the time of these actions what their total outcomes will be. Some consequences will conditionally depend on what others do.

It is precisely this indeterminacy and interdependence, both later explored in more detail, that is threatening MOAC's fulfillment of MH – and thus pose an *existential threat* to act-consequentialist theories. For this form of mutual interdependencies can lead to unfortunate situations in which particular combinations of actions lead to morally suboptimal or even disastrous results, while it seems that for any one of these actions acting differently would have made no difference for the better. It proves helpful to have a name for cases that contain such combinations of actions:

---

<sup>9</sup>I am convinced that this is precisely what John Donne had in mind when he wrote (cf. Donne [1923]): »No man is an island, Entire of itself. Each is a piece of the continent, A part of the main.«

**Definition 1.1 (TROUBLEMAKERS (tentative))** *A collective decision situation is a TROUBLEMAKER if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

**COLLECTIVE SUBOPTIMALITY** *together they would produce a morally suboptimal outcome and*

**INDIVIDUAL OPTIMALITY** *none of them could make a difference for the morally better by unilaterally acting differently.*

It must be pointed out that COLLECTIVE SUBOPTIMALITY is in no way meant to imply a collective action in the sense of a coordinated, joint action. That the agents »together produce« a suboptimal outcome through their actions merely states that the mere combination of the corresponding actions results in precisely these consequences, completely independent from shared intentions, joint decision-making procedures, or implicit coordination with each other. In fact, the absence of such a group-agent-like character generally makes TROUBLEMAKERS such TROUBLEMAKERS. This is because the path to the right kind of coordination as a way out of the CHALLENGE is »blocked«.

TROUBLEMAKERS raise a challenge as they apparently unveil an *inconsistency* between MOCOR and COLLECTIVELY MAXIMIZING. Assume that some agents find themselves in some TROUBLEMAKER scenario and that the agents perform a troublesome combination. According to COLLECTIVE SUBOPTIMALITY, they then produce suboptimal outcomes. Thus, according to COLLECTIVELY MAXIMIZING, at least one of them must have done wrong. However, given INDIVIDUAL OPTIMALITY, MOCOR seemingly entails that all of them did right. Thus, the CHALLENGE reveals a

serious material tension between MOCOR and COLLECTIVELY MAXIMIZING, two principles that are meant to explicate the same basic conviction, viz. CONGRUENCE. The CHALLENGE (as CHALLENGE<sub>int</sub>) then is apparently this: If

- MOAC theories are characterized by being theories that embrace MOCOR,
- and if the champions of MOAC are committed to accepting MH,
- and if the existence of TROUBLEMAKERS proves that MOCOR violates COLLECTIVELY MAXIMIZING,

then *MOAC theories are inadequate according to their very own standards.*

In this sense, then, as Shelly Kagan has put it, MOAC seems »to fail even by its own lights« (Kagan 2011, p. 108). Here is a tentative proposal for the CHALLENGE in this theory-specific sense (the CHALLENGE<sub>int</sub>) in terms of a tabular argument:

**Argument:** The ARGUMENT (tentative)

*P<sub>3T</sub>:* There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

*P<sub>MOCOR</sub>*: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

*P<sub>MH</sub>*: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

---

*C<sub>¬Adeq</sub>*: MOAC is not an adequate moral theory.

This variant of the CHALLENGE is much more than a philosophical puzzle for consequentialists. It is meant to reveal a truly serious inner-theoretical inconsistency, a matter of life or death for one of the most influential and traditional families of moral theories.

This book aims to show that the situation is not as hopeless as it might seem at this point. I claim that, contrary to the prevailing consensus, MOAC *has* the conceptual space to address the CHALLENGE without heavy, origin-denying modifications that other approaches suggested, and without having to abandon important theoretical or motivating ground. The goal of my project is to show that MOAC can, somewhat surprisingly, but without betraying its origins and core motivation, master the CHALLENGE.

## 1.2 Structure

This thesis has two parts, each of which has different objectives. The first part is reconstructive and preparatory in nature. The second part is constructive and serves to develop and defend my own approach.

The first part consists of three chapters. After giving an overview of the first part, I make some meta-comments on how I selected the relevant references and introduce some preliminary remarks on individual decision situations and consequentialist theorizing in Chapter [2] including useful notions and formalisms. Then, in Chapter [3], I sort through different strands of the debate on the CHALLENGE. In doing so, I differentiate between three different variants of the CHALLENGE – the CHALLENGE<sub>int</sub>, the NO-DIFFERENCE CHALLENGE, and the TRILEMMA – and explain how they relate to each other. In particular, I argue that ›the‹ strongest version of the CHALLENGE is the PYRAMID, a three-layered conglomerate of these variants. Along the way, I distinguish different kinds of TROUBLEMAKERS and collect some examples that will accompany us throughout this project. I conclude the part with Chapter [4]. In doing so, I primarily define how satisfactory approaches to CHALLENGE would look. I do this structurally by characterizing so-called solution spaces, which describe the space of possible ›theoretical movements‹ that can lead out of the CHALLENGE. But I also do it substantially by introducing (respectively collecting) and justifying a set of criteria. In this context, I also justify excluding a specific type of TROUBLEMAKERS, so-called CUMULATIVE EFFECTS CASES, from the scope of my project. Afterward, I will briefly discuss a rough taxonomy of existing approaches and give, with respect to three particularly prominent or, in the context of this project, ex-

citing approaches, some reasons why they are *not* satisfactory. This should suffice to motivate the development of my own approach.

The second, constructive part comprises five chapters. First, after a short overview of the second part, in Chapter 5, I introduce some more preliminaries that allow us to think and talk more precisely about collective decision situations and the assessments of MOAC theories therein. Next, in chapter Chapter 6, I problematize the common understanding of the CHALLENGE<sub>int</sub> as sketched above in Section 1.1. While it can be established that the ARGUMENT is not even valid, this comes at the cost of accepting that the CHALLENGE is rather a symptom of a deeper, conceptual issue. As a result, I present what I call the REAL CHALLENGE. The mastering of REAL CHALLENGE – in a way that does not reiterate the CHALLENGE – is the central goal of the remaining chapters. In chapter Chapter 7, I then advocate the APPROACH, a new conceptual framework that introduces a new kind of consequences to the objective consequentialist workbench. Together with one of various *prima facie* promising so-called *collective amendments* MOAC can solve the REAL CHALLENGE. In chapter Chapter 8, I discuss how camp MOAC should decide between various such amendments in a way that allows solving the REAL CHALLENGE but without violating PMH. Finally, I arrive at a concrete recommendation, a generalization of MOAC that I call MULTI-AGENT CONSEQUENTIALISM (short MAC).

In the concluding chapter Chapter 9, I summarize my work and dare a brief outlook on possible further work related to my results. I conclude by noting that collective contexts, hitherto a major Achilles heel of camp MOAC have now become a strength, a tangible competitive advantage in the constant battle for the best, most advanced moral theory.

## **Part I**

# **The CHALLENGE**

## **(and How *Not* to Solve It)**



# Overview of Part 1

Consequentialism has been historically confronted by a plethora of philosophical challenges. The CHALLENGE is a particularly crucial one – an issue that has persisted in academic discourses and catalyzed significant debates. This thesis positions itself within this ongoing discourse, aspiring to bolster the consequentialist framework against the CHALLENGE and related challenges that have their origin in collective contexts. For this, it is immensely important to first truly understand the CHALLENGE in all its blurriness and to elaborate it in its various modes from the depths of philosophical discourse. Of course, this also includes taking a critical look at existing approaches. Only what you can talk about properly, you can hope to solve; and only those who have screened the giants can choose the shoulders on which they want to stand.

Establishing these foundations is the goal of the first part of my project. It paves the way for the subsequent subproject of this thesis, setting the stage for a rigorous reinterpretation of the CHALLENGE and, finally, a new approach to the CHALLENGE. This part comprises three chapters. In Chapter 2, I lay some groundwork. The main goal is to start with a proper understanding of *individual* decision situations and a basic understanding of consequentialist theories (and their defining components) so that one can then move to *collective* decision situations.

*tive* decision situations. Next, in the extensive Chapter [3], the CHALLENGE is dissected through a reconstruction of several variants that can be reconstructed from different strands of the existing literature. The culmination of this analysis is the creation of a three-layered composite of different variants of the CHALLENGE, which I term the PYRAMID, and which is designated as the ultimate focal point of this project. Subsequently, adequacy criteria are established in Chapter [4], which serve as benchmarks for successful approaches to the CHALLENGE. Additionally, this chapter explores the shortcomings of at least three prominent approaches. The second part of the thesis is then committed to constructing my proposed resolution.

# Chapter 2

## Preliminaries I

In this section, I will introduce a set of concepts, notions, abbreviations, and naming conventions to ensure clarity and conciseness. Moreover, I will construct a semi-formal model that encapsulates (individual) decision situations as pertinent to (especially consequentialist) moral theories. This foundation will be instrumental in effectively communicating the vital properties and concepts that will be explored in the chapters that follow. While most of the initial portion of the thesis does not necessitate an understanding of these preliminaries, they become increasingly important as the discussion progresses.

### 2.1 Individual Decision Situations

Fortunately, for a sufficient degree of precision, we do not need something as complex as a complete semantic for advanced deontic logic. The relatively simple framework, built upon the following key building blocks, does suffice:

**Agents, Decision Situations, Options:** Every now and then, an *agent*<sup>10</sup> finds herself in a *decision situation*, i.e., a situation where she has to choose between several *options*.

---

<sup>10</sup>By default, we may think of natural persons as agents, but there might also be group agents (List and Pettit 2011). In the context of this work, the important distinction is between cases where one individual entity makes a decision and performs an action (individual decision situations) and cases where several such entities are involved (collective decision situations). It is not important whether these agents are human persons or, say, corporations. We might replace Ann and Ben with two executive boards of companies *A* and *B*.

**Actions:** Each option has a *corresponding* action.<sup>11</sup> Options can be *instantiated* by *performing* the *action* corresponding to that option. Under normal conditions, an agent performs the action corresponding to the option they chose. Following an unwritten rule of normative ethics, I use » $\Phi$ « (and sometimes  $\psi$ ) for denoting options and actions. The context should make clear whether the one, the other, or both are meant.

**Contexts and Consequences:** Performing an action has *consequences*. An action's consequences are what would be the case if that action were performed but what would not otherwise be the case. Strictly speaking, options have consequences (only) in a derived sense, namely the consequences that their corresponding actions will or would have (again, relative to some context). That said, I will generally choose the less long-winded way of expression and not mention this every time.

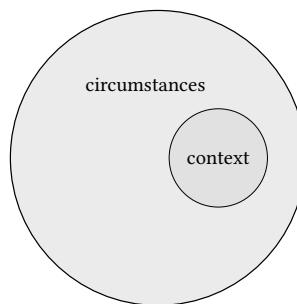
The (actual) consequences of an action (usually) depend on a number of facts that obtain (at the moment of the action). We call those facts on which the (actual) consequences of the action depend, the *actual context* of a decision situation (at a point in time). We call the totality of facts (at a point in time) the circumstances. Figure 2.1 visualizes the relation between circumstances and context, while Figure 2.2 visualizes the relation between different kinds of (potentially relevant) contexts. Accordingly, the consequences of an action would have been different if the (actual) context had been different. We call contexts that could be the case *contingent contexts*. Contingent contexts can play a role in the

---

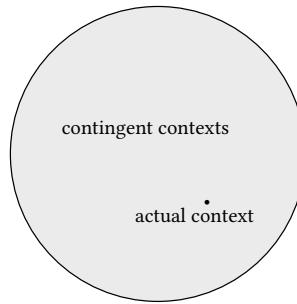
<sup>11</sup>Distinguishing between options and actions is essential for clarity. An agent faces *several* options but performs *one* action – or *at most one* if inaction is not necessarily counted as an ›act of omission‹. Options are potential ways to act, while an action is the enactment of a chosen option, manifesting an event in the world.

normative assessments of actions, namely when not the actual context but, for example, the contexts that can be expected for the agent are considered normatively *relevant* (more on this later). For the sake of simplicity (and because it corresponds to the practice in the debate), we will generally assume that contexts are static relative to decision situations, i.e. they do not change during the decision situation. (However, this assumption cannot be maintained over the course of this project and will be relaxed when the time comes.)

While for this project it is critical to describe decision situations with quite some formal precision, their metaphysical and ontological nature is not. Accordingly, I focus on establishing a clear language and framework for analyzing and modeling them and refrain from engaging in metaphysical discussions. I would be willing to be carried away into asserting that decision situations, in the context relevant to this thesis, can be conceptualized as some kind of structure involving (transient) states of affairs<sup>[12]</sup> involving the above specific elements as building blocks.<sup>[13]</sup> It's worth noting that decision situations may



**Figure 2.1:** Circumstances and context. The actual (or a contingent) context is a part of the actual (or contingent) circumstances.



**Figure 2.2:** Set of potentially (normatively) relevant contexts. The actual context is one potentially relevant context, all other potentially relevant contexts are contingent.

<sup>[12]</sup>A state of affairs is *transient* if, and only if, it obtains at one time and not another (cf. Textor 2021).

<sup>[13]</sup>One may think here of branching time models in the sense of A.N. Prior (Arthur N. Prior 1967; Arthur N. Prior 1955), in which options are possibilities to choose between potential futures that are ultimately overlapping sets of different kinds of propositions or states of affairs. Stit semantics seem to me to represent a particularly promising modeling of this idea (cf. Belnap, Perloff, and Xu 2001; Harty and Belnap 1995; Harty 2001).

encompass additional elements, and in different contexts, they may be characterized differently. Further, one should be careful not to confuse the representation of a thing with the thing itself: The formal descriptions employed in this thesis are, in a sense, models representing decision situations, typically merely hypothetical ones. Sometimes this differentiation matters, often it does not. Where I think that it does matter, I try to be as explicit as possible.

With this conceptual groundwork laid, I suggest the following definition:

**Definition 2.1 (Individual Decision Situation)** *An individual decision situation is a situation in which a single agent is presented with multiple options, and, within a given context, each action that corresponds to an agent's options has an associated consequence.*

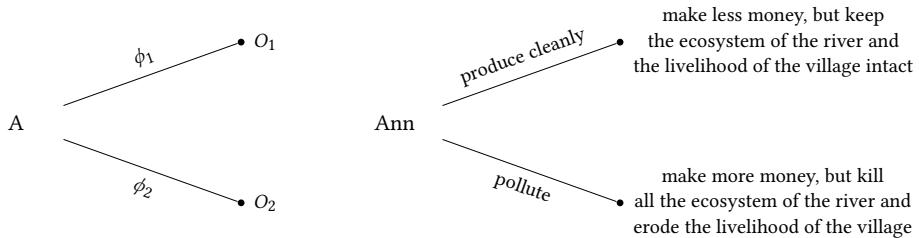
Here is a description of an individual decision situation, a modified version of an example by Felix Pinkert (Pinkert [2015]), which will later take a central role in this project:

**Case 2.1 (FACTORY)** *Ann owns a factory near the river. Ann can either produce cleanly or pollute. Polluting would allow Ann to produce significantly cheaper and, as a result, make some extra cash through which she can afford another week of vacation at a diving resort. But polluting would also kill all the fish in the river and erode the livelihood of a village downstream.*

Before I proceed, a word on the relationship between cases and decision situations: cases are best understood as being *descriptions* of decision situations. This means that we can easily jump between some case (description), which is a linguistic entity, and the decision situation described by it. Of course, this presupposes that the description is such that it actually describes a decision situation, i.e., that the description is coherent, consistent, and suf-

ficiently complete. In particular, such a description must specify a decision situation in terms of the components outlined above and with regard to a relevant context. A proper case in that sense insofar warrants the *existence* of the described situation in a metaphysically ›lightweight‹, deflationary way: if there is a coherent, consistent, and sufficiently complete case description, then there is a corresponding decision situation (together with a relevant context) that is described. This goes well with the arguably very convincing view that not only actual but also conceptually possible<sup>14</sup> decision situations – let us call them »thought experiments« – are relevant to and a benchmark for normative theories. This view certainly corresponds to the practice of normative ethics (and rationality theory)<sup>15</sup>

Let us now furnish our toolbox with some representational devices. Sometimes it is handy to depict a decision situation – or rather its structure – visually. We can depict individual situations of some agent *A* with some options, say,  $\phi_1$  and  $\phi_2$  with consequences  $O_1$  and  $O_2$  in simple »decision trees«. Here is a generic example on the left and an instance for FACTORY on the right:



<sup>14</sup>The reader might feel free to weaken the claim to nomologically possible situations. We just could add to the list of conditions above – being coherent, consistent, and sufficiently complete – the condition of being in line with the laws of nature.

<sup>15</sup>Should this be doubted, I would like to refer to the myriad of trolley examples and many other thought experiments that not only populate introductory courses but are also commonly elevated to the standard of correct moral assessments.

Obviously, we can generalize this device to decision situations with an arbitrary number of options.

Having some formal notations at hand pays off later. Given a specific decision situation  $D$  of an agent  $A$  with options  $\phi_1$  to  $\phi_n$  I call  $\Phi = \{\phi_1, \dots, \phi_n\}$  the *option space* of  $A$ .<sup>16</sup> I assume that a decision situation implies that a *choice* can be made, that is, in particular, that every agent always has at least two options, i.e.,  $|\Phi| \geq 2$ . I want to be able to represent dynamics and uncertainty later on. To this end, I allow more than one context per decision situation by defining a relevant set of contexts relevant contexts  $C$  for a given situation  $D$ . What makes a context *relevant* is a question that demands an answer from a normative theory – I postpone it to the next subsection. Furthermore, we call  $\mathcal{O} = \{O_1, \dots, O_m\}$  the set of consequences of  $D$ . As we assume that the outcomes are a function of action and context, we get  $m = k \cdot n$  relevant outcome, where  $k$  is the number of relevant contexts, i.e.,  $k := |C|$ , and  $n$  is the number of options. Accordingly, we can model this relationship between actions and contexts on the one side and the consequences on the other as an *outcome function*  $\text{Out} : \Phi \times C \rightarrow \mathcal{O}$ . For singletons  $C = \{C\}$ , we can rewrite for convenience:  $\text{Out} : \Phi \times \{C\} \rightarrow \mathcal{O}$  to  $\text{Out}_C : \Phi \rightarrow \mathcal{O}$ .

When more than one decision situation is under consideration, I sometimes use indices for disambiguation, e.g., I write  $C_D$  for the set of relevant contexts of a decision situation  $D$  or  $\Phi_D$  for the option space of that decision situation; where there's no need, I leave them out. It is practical and convenient to have a common agreement on formal entities representing a decision situation as well (and not only of their constituents introduced above): Let  $D$  be an individual decision situation of agent  $A$  with options  $\Phi$ , a corresponding

---

<sup>16</sup>I restrict my investigation to cases with finitely many options.

outcome function  $\text{Out}$ , and a set of relevant contexts  $\mathcal{C}$ .  $D$  is represented by the tuple  $D := \langle A, \Phi, \text{Out} : \Phi \times \mathcal{C} \rightarrow \mathcal{O} \rangle$ <sup>17</sup> Often, we can refer to a decision situation  $D$  with a set of relevant contexts  $\mathcal{C}$  just by giving one such tuple and, thereby, implicitly introducing the  $\mathcal{C}$  and also the set of outcomes  $\mathcal{O}$  to the discourse through explicating only the signature of the outcome function.<sup>18</sup>

Finally, let  $I$  denote the set of all individual<sup>19</sup> decision situations.

After these meta-comments and preliminary formalities, it is now time to start the reconstructive work.

## 2.2 Moral Theories, the Rightness Predicate

Now that there is a clear understanding of individual decision situations, we can turn to the normative questions to be asked about these situations. In particular, with respect to MH and COLLECTIVELY MAXIMIZING, we next need a precise way of writing about the rightness of options or actions according to some moral theory. Most fundamentally speaking, moral theories are principled answers to a specific question, viz.

(N) What is right for an agent to do (and why)?

Usually, this question is answered in a principled way by moral theories with regard to individual decision situations. Now that we have a sufficiently clear conceptual framework of decision situations at hand, we can turn to the substantial, normative aspects of (N).

---

<sup>17</sup>Note that this is meant to establish an implicit naming convention: If we are considering two individual decision situations  $D_1$  and  $D_2$  we can easily switch to the level of formal representation by switching to the corresponding tuples  $D_1$  and  $D_2$ .

<sup>18</sup>The signature of a function is the definition of the inputs and outputs of a function. For instance, » $\text{Out} : \Phi \times \mathcal{C} \rightarrow \mathcal{O}$ « is the signature of the outcome function  $\text{Out}$  that maps a pair consisting of an option from  $\Phi$  and a context from  $\mathcal{C}$  to a set of  $\mathcal{O}$ .

<sup>19</sup>I will introduce *collective* decision situations later in a similar way, later in this book. For now, individual decision situations will do.

Before I turn to the family of moral theories this project aims at rescuing from the CHALLENGE, i.e., MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (MOAC), I briefly introduce some normative terminology and a more formal way of writing precisely about moral assessments relative to some theory. A wide variety of interrelated normative concepts are relevant in the context of (N). For instance, how does what is the right thing to do relate to what one ought to do? Or to what is wrong to do? And what does it mean for some outcome to be good or bad? These and related questions have at least two dimensions. First, they can be about different kinds of *normative statuses* (or properties); second, they can be about the different things that can have these statuses and properties: agents, options, actions, intentions and motives, states of affairs, ... In everyday language, the uses of the corresponding terms are not precisely fixed. In the context of moral philosophical theorizing, the challenge is to find, on the one hand, theoretically fruitful characterizations and explications, but, on the other hand, characterizations and explications that are also sufficiently close to the common usage. Otherwise, our moral philosophical theories might cease to be useful for our everyday discourses. It is thus a challenge typical for analytic philosophy.

The following characterizations are intended to conform to the ideal of analytic philosophy insofar as I make my understanding explicit. I do not, however, claim my characterizations to be fundamentally, metaphysically true (regardless of whether it can make sense at all to make such claim, which I generally doubt). Instead, the picture drawn below is intended to be sufficiently close to that of the practice of the philosophical discourses relevant to this project, while at the same time proving useful for my own purposes.

I propose to understand the concept of the normative (or, in the context of this thesis, moral) status to be non-monolithic. We should distinguish between *deontic* and the *evaluative* statuses. In this book, I am concerned with the deontic status of options and actions and thus with whether, for example, a particular action of an agent in a particular decision situation is right, wrong, neutral, forbidden, permitted, or obligatory<sup>20</sup>. Furthermore, I limit myself to questions regarding rightness or wrongness since this corresponds most closely to the literature on the CHALLENGE, and there does not seem to be any unfair simplification of the matter as a result of this limitation. I do not exclude the possibility that there is a fundamental difference between rather *prescriptively* loaded notions like obligatoriness, permissibility, and forbiddenness on the one side and rather *descriptive* sounding<sup>21</sup> notions like rightness and wrongness.

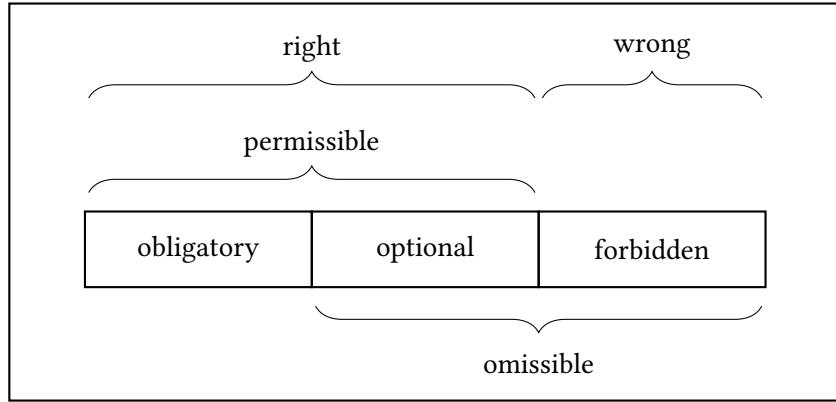
The exact relation between these properties need not concern us at this moment, but here is a proposal based on a suggestion by Krister Bykvist (Bykvist 2003, pp. 45):

- (1) An action  $\phi$  is *obligatory* for an agent  $A$  in (a decision situation)  $D$  if and only if  $\phi$ 's outcome would be better than the outcome of any alternative action for  $A$  in  $D$ .

---

<sup>20</sup>I believe that understanding actions respectively options as the primary bearers of deontic statuses correspond to our everyday linguistic practices. Especially in deontic logic, it is typical to use propositions or states of affairs as carriers instead. Accordingly,  $Op$  for a deontic operator  $O \dots$  then reads approximately »It ought to be the case that  $p$ «. While I do not want to deny that there is a meaningful way to talk like that, there is some evidence that some of the classical ›dilemmas‹ of deontic logic result from not referring to agency and actions instead, as stit semantics do (cf. most importantly Harty 2001). For this project, I settle on actions and options as primary, or at any rate essential, bearers of deontic statuses.

<sup>21</sup>Whether they are actually descriptive depends, obviously, on certain meta-ethical assumptions, first and foremost, moral realism. However, I do not commit to any such view here as it does not matter with respect to the CHALLENGE.



**Figure 2.3:** The relation between the different kinds of deontic statuses according to the CONSEQUENTIALIST STANDARD VIEW (cf. McNamara and Van De Putte 2022 Fig. 3).

- (2) An action  $\phi$  is *right* for an agent  $A$  in (a decision situation)  $D$  if and only if  $\phi$ 's outcome would not be worse than the outcome of any alternative action for  $A$  in  $D$ .
- (3) An action  $\phi$  is *wrong* for an agent  $A$  in (a decision situation)  $D$  if and only if  $\phi$  is not right for  $A$  in  $D$ .

Relevant here is not whether these statements are substantially correct. What is at issue is that these quotes emphasize that for champions of MOAC, typically, the scheme shown in Figure 2.3 applies: if there is only one right action, it is also obligatory (and, if there are several right actions, then it is obligatory to perform one of these); and the set of right (wrong) and the set of permitted (forbidden) acts are extensional equivalent. I call this the CONSEQUENTIALIST STANDARD VIEW. Thus, where necessary or useful, I feel free to switch unhesitatingly between describing an action as being right and describing it as being permitted on the one hand and describing an action as being wrong and describing it as being forbidden on the other, especially when it comes to the interpretation of specific quotations.

Like the CONSEQUENTIALIST STANDARD VIEW, I, for the most part, ignore the possibility of morally neutral or genuinely conditional statuses. This is indeed quite a common view. For instance, note that Bykvist's definitions just quoted even exclude such statuses. Neutral actions, i.e., actions that are neither right nor wrong, are excluded because all actions that are not right are, according to (3), wrong. For similar reasons, actions with *genuinely conditional* deontic status are excluded, i.e., actions that are right (or wrong) only if a certain condition holds, but neither wrong nor right independent from whether said condition holds or not. Either such an action is wrong or it is not wrong. In the second case, that action is, again, right according to (3). Thus, in both cases, that action would have a non-conditional moral status and, hence they would no longer have *genuinely conditional* deontic status. A contradiction. However, we will see that consequentialists are well advised to make conceptual space for at least genuine conditional deontic statuses, at least in their current form of the objective consequentialist framework. (The version of consequentialism that I propose in the context of this book does not need conceptual space for such exotic and esoteric statuses. But I don't want to get too far ahead of myself.)

Furthermore, there are so-called *evaluative* statuses, of which there are two kinds, namely *absolute evaluative statuses* – being good, being bad, and possibly being (evaluative) neutral – and the *comparative evaluative statuses* – being better than, being worse than, and being equally good as. The latter statuses are always relational, i.e., relative to other options or actions. The following distinction proves useful for this project: Theories that ascribe evaluative statuses are called *axiologies*; theories that ascribe (moral) deontic

status are *moral theories*.<sup>22</sup> I will return to axiologies and their role within consequentialist theories later in this chapter.

We can now formally introduce moral theories as theories that – quite in the spirit of question (N) above – operate on individual decision situations and assign the options within their corresponding option space a deontic status in a principled way. This exclusive focus on individual decision situations may seem ›innocent‹ but against the backdrop of the collective character of the CHALLENGE (i.e., of TROUBLEMAKERS), this commonly accepted aspect of moral theories is significant for this project. It is thus worth to be a bit more explicit in this regard. In this sense, I take it that the following principle<sup>23</sup> expresses a widely accepted but seldom explicated view:

### **Principle 2.1 (METHODOLOGICAL INDIVIDUALISM)**

*The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the act of a moral agent has this status merely in a derivative sense, that is, that deontic status is a function of the deontic status of certain options of moral agents.*

---

<sup>22</sup>Of course, there are also completely different conceptions of what a moral theory is and how moral theories relate to value theories and possibly other sub-theories. For example, Marc Timmons suggests that moral theories are composed of a »Theory of Right Conduct« and a »Theory of Value«, which makes statements about both »Intrinsic Value« and »Moral Worth«, cf. Timmons [2001] pp. 9. For the present (consequentialist) project, however, a theory of intrinsic value (i.e., an axiology) is sufficient – and what Timmons calls a »Theory of Right Conduct« is, in my picture, a certain part of the moral theory that is typically characterized by a specific criterion of rightness (plus some relationships between further moral statuses, see above). But we will come to the details in a moment. What is important here is this: By no means do I want to offer a general analysis of the conception of moral theory. I do not want to commit myself to *any* one such conception. (Actually, I do not believe that it is worth to search for any *one* such conception, since different conceptions and conceptualizations are differently well or poorly suited for different purposes.)

<sup>23</sup>Originally, I thought I had come up with a wonderfully apt name for this principle. But it turns out that there is an established principle of this name in the social sciences (cf. Schumpeter [1908]; Weber, Roth, and Wittich [1978]; Heath [2020]) that actually has a close conceptual relationship to the principle proposed here as it calls for social phenomena to be explained by appeal to individual actions.

METHODOLOGICAL INDIVIDUALISM may raise various questions, the most pressing of which is arguably why options and not actions should be the primary bearers of deontic status. Although I believe that, in principle, both views – that options or actions are the primary carriers – can be defended, I also believe that the question of the *appropriate* explication of the principle does not ultimately depend on some metaphysical truth, but on which perspective one adopts with which particular theoretic interest. In other words, the choice of the specific conceptual framework should ultimately also depend on its usefulness for the respective endeavor of theory building. If one aims to describe the observable world ›from the outside‹, then it is probably expedient to regard actions as the primary bearer of deontic status because options are, in a sense, not observable, are no events in the world. However, if one is interested in understanding decision situations and wants to build and develop theories about what the right thing to do for an agent is in a given decision situation (cf. question (N) above), then it is options that arguably should be considered as such primary bearers. Actions (as the manifestations or realizations of options) then inherit their moral status directly from the options that correspond to them. If an action is right (or wrong), it is *because* the agent has realized a right (or wrong) option through this very action. Since this project indeed revolves around decision situations and focuses on what is right for an agent to do in that very sense, it is essential to know which option is right and, hence, which action is right to perform. Thus, I suggest the version of METHODOLOGICAL INDIVIDUALISM given above be adequate in this project's context.

Most importantly in the context of the CHALLENGE, METHODOLOGICAL INDIVIDUALISM tells us something relevant about combinations of

actions: If a combination of actions has deontic status, then, arguably (in the absence of better candidates), this status is a function of the moral status of its parts, i.e., the individual actions that make up that combination. Note that one can embrace **METHODOLOGICAL INDIVIDUALISM** without committing to the view that combinations of actions can have (or even to the view that they normally have) deontic status at all. All it suggests is that *if* some combinations have such status, then they have in virtue of the deontic status of the individual actions that constitute these combinations. A simple principle in line with **METHODOLOGICAL INDIVIDUALISM** could be, for instance, that a combination of actions is right, if all actions it contains are right; and that such combination is wrong if all contained actions are wrong (we discuss such principles at the very end of this part in the context of Frank Jackson's approach to the **CHALLENGE**, cf. Jackson [1987]). Further, **METHODOLOGICAL INDIVIDUALISM** does not rule out the existence of group agents nor that their actions have some kind of genuine deontic status. As long as they qualify as genuine agents (cf. List and Pettit [2011]; Pettit and Schweikard [2006] for an elaborated account), they *are* agents and, as such, can populate individual decision situations. However, **METHODOLOGICAL INDIVIDUALISM** rules out that *mere* collectives, loose groups that *do not qualify as agents* in any substantial sense, can be the ultimate addressees of moral theories' guidance. If some combination of actions is made up of actions of individuals standing in such loose connection and if that combination has some deontic status at all, then this deontic status is a function of the deontic status of the individual agents' actions. According to **METHODOLOGICAL INDIVIDUALISM**, thus, a moral theory is not allowed to stop or even start with the assignment of a deontic status to such combinations.

Since one could assume that not every moral theory has to be applicable to every decision situation, it can be useful to restrict one's reasoning about a moral theory  $T$  to a set  $I_T$ , the set of decision situations for which  $T$  has something to say, i.e., to  $T$ 's *domain*.

It will come in handy later to have a shorthand for »given a decision situation  $D$  (of agent  $A$  with options  $\phi$ ) it is right (for  $A$ ) to  $\phi$  given context  $C$  according to moral theory  $T$ «. For this, we will use a rightness predicate  $R$  and write

$$T, D, C \vDash R\phi.$$

We can give a set-theoretic semantic for  $R$ . For this, we first define some useful sets. Let  $D' \subseteq I$  be a set of individual decision situations.  $C_{D'}$  denotes the set of all the relevant contexts of the decision situations in  $D'$  and  $\Phi_{D'}$  denotes the set of all the options available in the decision situations in  $D'$ , i.e.,

$$C_{D'} := \bigcup_{D \in D'} C_D, \quad \Phi_{D'} := \bigcup_{D \in D'} \Phi_D.$$

We can think of a moral theory  $T$ , then, in terms of a function  $T$  that maps a decision situation  $D$  from  $T$ 's domain  $I_T$  with an option space  $\Phi_D$  and fitting context  $C \in C_D$  to a set of right actions  $\Phi_T \subseteq \Phi_D$ , i.e.,

$$T : I_T \times C_{I_T} \rightarrow \Phi_{I_T}$$

with

$$T(D, C) := T(D, C) = T(\langle A, \Phi, \text{Out} : \Phi \times C_D \rightarrow \mathcal{O} \rangle, C \in C_D) \subseteq \Phi$$

which is to be interpreted as

$$T(D, C) = \{ \phi \in \Phi_D \mid \phi \text{ is right for } A_D \text{ given } C \text{ according to } T \}.$$

Note that the function  $T$  is undefined for pairs of decision situations outside of  $T$ 's domain or irrelevant contexts.

We can now semantically define the shorter, but equivalent definition

$$T, D, C \vDash R\phi \text{ if and only if } \phi \in T(D, C).$$

For now, we can leave open whether there is anything more to say about a wrongness predicate  $W$ ... than that it states that an action is not right, i.e., whether we should accept

$$T, D, C \vDash W\phi \text{ if and only if } T, D, C \not\vDash R\phi.$$

and thus whether we can simply define

$$T, D, C \vDash W\phi \text{ if and only if } \phi \notin T(D, C).$$

If the standard view is true, there is nothing to be said against this simple formalism (though later, in the context of collective contexts, I will raise doubts about this very assumption).

At this point, it might seem as if the definition of a predicate of rightness is all that constitutes a moral theory. While this is right in a certain sense, such a predicate of rightness involves much more than just a *criterion* of rightness. What may sound a bit cryptic at first will be explained in the next section in much more detail. For the time being, however, it is sufficient to remember the criterion of rightness, which has already been roughly outlined in the introduction and which, so to speak, plays a main role in this work. Recall

**Criterion 1.1 (MOCoR (tentative))** *An action is right if and only if there is no alternative action that would actually lead to better consequences.*

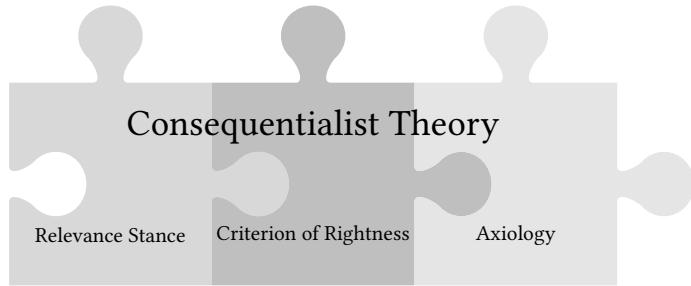
We can capture this a bit more precisely as

**Principle 2.2 (MOC or)** *Let  $D$  be an individual decision situation involving an agent  $A$  with a set of options  $\Phi$  and with actual context  $C$ . An action  $\phi \in \Phi$  is right for  $A$  if and only if, relative to  $C$ , there is no alternative action  $\phi' \in \Phi$  with better consequences than  $\phi$ .*

As already mentioned in the introduction, any moral theory embracing this criterion of rightness is, by definition, part of the family of theories I call **MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (MOAC)**. But what is the specific function that makes up  $T_{\text{MOC or}}(D, C)$ ? I cannot give a formal condition for it at the moment, because the framework developed so far is missing something essential: a way to morally evaluate the consequences of actions and to compare them with each other. In the next section, which concludes this chapter, this gap will be closed.

## 2.3 MOAC and Its Three Modules

Looking back, we find at least three potential aspects a completely specified criterion of rightness could operate on: the agent, the options, and the consequences. These three aspects correspond to the traditional triad of ›pure‹ families of moral theories: The family of virtue ethics focuses on the agent, especially their character; for the family of deontological theories, the moral status of options and actions is, first and foremost, a function of the action itself, its type and properties, i.e., its universalizability and adherence to certain (universalizable) rules; for the final family, consequentialism, the moral status is determined *solely* by the consequences of the options. Naturally, there can be hybrid theories. While admittedly painted with a broad brush, this distinction suffices for demarcation purposes.



**Figure 2.4:** The modular nature of consequentialist theories. They consist of a view regarding which contexts are relevant, a criterion of rightness, and an axiology. MOAC consists of OBJECTIVISM, MAXIMIZATION, and an axiology that allows ranking outcomes (as discussed in the next section).

Since consequentialist theories know no limitations as to their applicability, their domain is the set of all decision situations.<sup>24</sup> It is thus appropriate to characterize consequentialist theories as exactly those moral theories that embrace a particular fundamental, meta-normative principle, viz.

**Principle 2.3 (ESSENCE OF CONSEQUENTIALISM)** *For every individual decision situation and an associated set of relevant contexts: the moral status of an action is solely determined by the morally relevant qualities of the consequences of the action (typically relative to the morally relevant qualities of the consequences of its alternatives).*

Looking more closely at this condition, I suggest conceptualizing full-fledged consequentialist theories as consisting of three modules (cf. Figure 2.4): First, they need a view concerning which contexts are relevant (and thus which consequences of an action count, e.g., the actual ones or the expected ones, etc.) which I call a »relevance stance«; second, they need a *criterion of rightness* that determines the deontic status of options and actions based on

---

<sup>24</sup>This statement should be taken with a grain of salt. I am firmly convinced that this is indeed the default view of the vast majority of consequentialists, even if it is never actually made explicit (at least, I am not aware of any explicit statement). Moreover, one can possibly raise doubts about this view as I argue later, in the last chapter of this part. In fact, the question is relevant for this project, because I will motivationally put forward this self-image of consequentialism as a ›deontic complete‹ theory later on.

the moral quality of their consequences; and, finally, a background theory known as *axiology* that determines what moral qualities (or value) these consequences have. In other words, axiologies put moral value on the workbench of consequentialists, the criterion of rightness defines how these values affect the normative status of options and actions in their consequences, and the relevance stance determines which consequences are relevant in this regard.

The assertion that consequentialist theories are fundamentally defined by three modules, with the criterion of rightness being just one of them, seems to conflict with earlier statements in this book. Up until now, I maintained that a moral theory falls under MOAC if and only if it adheres solely to *one specific* criterion of rightness, namely MOCoR. Nevertheless, this is not an inconsistency but rather an ambiguity as MOCoR is, in fact, a composite of both a criterion of rightness (in the »module sense« as previously introduced) that I shall call »MAXIMIZATION« – which corresponds to the »M« in MOCoR – and a relevancy stance that I shall call »OBJECTIVE VIEW« – referred to by the »O« in MOCoR. The third module, however, the *axiological* background theory, is deliberately left unspecified because this is where the various MOAC theories diverge. Thus, MOAC is just one family of partially but not fully defined consequentialist theories.

A detailed and incremental introduction to these modules is invaluable for this project. This is particularly true concerning the axiological module because one specific part of the axiological background theories that concerns aggregational questions will play a central role in the later stages of my project (while the other part, concerning what has inherent value in the first place, does not matter at all for my purposes).

### 2.3.1 Relevance Stances

What I call »relevance stance« is a view regarding the question of which perspective on decision situations is relevant for their normative assessment. There are two kinds of relevant stances that are prominent in the consequentialist camp.

One is *objective*, in the sense that it looks at the situation from the outside. Intuitively but misleadingly formulated, this is supposed to express that the moral status of an action depends solely on what is the case. This statement is misleading because one can, of course, claim that the fact that an agent believes, for example, that something is also the case – after all, it is a fact. However, this is precisely what is *not* considered a relevant fact according to that view. Instead, the relevance of epistemic and doxastic aspects is meant to be discarded.<sup>[25]</sup>

To express the idea more precisely, we can thus resort to the concept of the actual context of a decision situation. We have introduced the *actual context* of a decision situation as that specific part of the situation's circumstances on which, given a particular action, its consequences depend (cf. page [22]). However, because the consequences of an action do not commonly<sup>[26]</sup> depend on the beliefs and desires, and more generally, do not depend on the mental

---

<sup>25</sup>For this reason, I personally prefer to frame the following distinction as one between non-epistemic and epistemic stances. However, at least calling the non-epistemic stance »objective« is rather the default in the consequentialist debate (cf. Railton [1984], Sinnott-Armstrong [2022]), and I will not unnecessarily divert from that convention.

<sup>26</sup>Exceptions are conceivable. Imagine that Timo, who has a serious headache, can either take a magic drug or not. If he takes the drug, it will bring relief to Timo if and only if he believes it to do so. In this case, the consequences of Timo's drug-taking depend objectively on Timo's beliefs. Similar dependencies of outcomes and an agent's mental states appear in cases like Kavka's toxin puzzle (Kavka [1983]). In such rather exotic cases, the agent's mental states (or the mental states of other persons) belong to the actual context.

life of the agent, they (normally<sup>27</sup>) do not belong to this context. We therefore define:

**View 2.1 (OBJECTIVE VIEW)** *The moral status of an action within a decision situation depends solely on the actual context of the decision situation.*

Under the OBJECTIVE VIEW, thus, the agent is only the ›executive organ‹ of the decision-making situation. It does not depend on their convictions, principles, attitudes, dispositions, nor on their particular (epistemic) history that led them to their situation at this point in time. It is entirely irrelevant what they have evidence for, what they are justified to believe, how strongly they are justified to believe, etc. Thus, according to the objective stance, each decision situation has only one *relevant* context, namely the *actual* context.

This is the objective relevance stance.

The second kind of stance is *subjective* insofar as these stances assign relevance to epistemic aspects. Accordingly, they are agent-relative. These stances, thus, make the moral status of an agent's actions also a function of, for instance, the agent's beliefs or what the agent ought to believe and so on. According to subjective stances, there is usually a multitude of relevant contexts, for instance, all the contexts that the agent actually considers possible or that the agent may consider possible according to some normative *epistemic* standard (here, it depends very much on the details of the respective subjective stance). Based on a distinction made by Michael J. Zimmerman (cf. 2014), one can, therefore, differentiate between at least the following two stances:

---

<sup>27</sup>Cf. Footnote 26. For the sake of better readability, I will omit this kind of hedging from now on.

**View 2.2 (SUBJECTIVE VIEW)** *The moral status of an action within a decision situation depends solely on the agent's beliefs about the possible contexts.*

**View 2.3 (PROSPECTIVE VIEW)** *The moral status of an action within a decision situation depends solely on the beliefs that a perfectly epistemically rational agent would have about the possible contexts given the epistemic history of the agent.*

We need not dive into the details of these views in the context of this project since we are primarily concerned with defending MOAC theories and, thus, by definition, objective consequentialist theories. Nevertheless, the PROSPECTIVE VIEW plays a role in the context of the thesis, on the one hand, in reconstructing the CHALLENGE and, on the other hand, because I want to argue in the second part of my thesis that MOAC adherents can learn something important from it.<sup>28</sup>

One can take two positions with respect to the validity of these general perspectives: either one can see it as a dispute concerning the question of which view is >really correct<. Then subjective/epistemic and objective/non-epistemic notions are *mutually exclusive*. Alternatively, one can see them as complementary, as two justified perspectives that together offer a richer moral framework. The complementary view can take several forms. For instance, one can disambiguate the same moral predicate subjectively and objectively (for instance, being right) or advocate an objective view for some predicates (say, being right) and a subjective view for others (say, being obligatory). According to the latter, one can hold that the question of what an agent *ought* aims at an answer that should be action-guiding and therefore formulated

---

<sup>28</sup>The SUBJECTIVE VIEW is not considered particularly plausible anyway, see for example Jackson [1991]. It has been included here only for contrastive purposes.

subjectively, while at the same time, one can hold an objective view for moral *rightness* (cf. Ord [2005]; Andrić [2013]). According to a first, one could, for example, hold that sometimes we are concerned with subjective rightness, while in other contexts, we are interested in objective rightness. Then, objective rightness arguably equals the ›epistemic limit case‹ of subjective rightness. Derek Parfit once expressed such a view (Parfit [1988], p.2): »if, when acting, we know all the relevant facts«, then the »two kinds of rightness [...] coincide«. Somewhat more precisely,<sup>29</sup> we can capture that idea like this:

**Principle 2.4 (EPISTEMIC LIMES)** *Let  $T_O$  be an adequate objective (i.e., non-epistemic) moral theory and let  $T_S$  be an adequate subjective (i.e., epistemic) moral theory. For a given decision situation involving an agent A with a set of options  $\Phi$  and a set of relevant contexts C: If A knows all relevant facts and has no incorrect relevant beliefs, then*

$$T_O, D, C \vDash R\phi \text{ if and only if } T_S, D, C \vDash R\phi.$$

In other words, according to EPISTEMIC LIMES, adequate objective and adequate subjective moral theories are extensionally equivalent under the assumption of perfect epistemic situatedness of the involved agents.

---

<sup>29</sup>However, the whole idea sketched here presupposes that knowing something implies absolute certainty and thus a credence of 1. Such an assumption might actually be implausibly strong and, thus, might be doubted for good reasons (cf. Williamson [2002], MacFarlane [2023]). Under such a ›softer‹ concept of knowledge, an agent could then very well know that an option  $\phi$  has the best consequences, and yet the expected value of  $\neg\phi$  could still be greater than the value of  $\phi$ . However, such concepts of knowledge can be set aside here: Where I bring the following principle (EPISTEMIC LIMES) into play later, I need it in order to be able to argue about certain parts of the literature in which knowledge is presupposed in the context of the discussion of the CHALLENGE. Either these contributions are to be read in such a way that knowledge *implies* certainty, or they are not. I write under the assumption of the first case and only then do I need EPISTEMIC LIMES. If a ›softer‹ concept of knowledge is meant, one in which knowledge does *not* imply credence of 1, my general considerations on uncertainty and underdetermination can be applied.

In the context of this work, we do not really need to commit ourselves in this regard. But where subjective and objective consequentialist theories play a role in the course of the following chapters, it makes sense to take a complementary perspective, according to which the objective view is an ideal case of the subjective one, i.e., a case in which the agent knows everything that is relevant and believes no falsehoods.

### 2.3.2 Criteria of Rightness

The second module is a criterion of rightness. This criterion defines which deontic statuses are assigned for options of the decision situations in the domain of the theory (and why they are assigned). In principle, this could be a criterion of what is right or wrong (or of what is forbidden, or what is permissible, ...), and then all other properties could be implicitly assigned according to specific relationships to each other (for example, according to the CONSEQUENTIALIST STANDARD VIEW outlined above). Without loss of generality, this project limits itself to the last type of criteria, if only because this corresponds with the consequentialist tradition, especially with an eye on the literature regarding the CHALLENGE. A criterion of rightness is both classificatory and explanatory. It is *classificatory* in the sense that it classifies an option as right or wrong; it is *explanatory* insofar that it explains why<sup>30</sup>

---

<sup>30</sup>The typical »if and only if« formulation is not to be read a material biconditional in such *explanatory* definitions. Such reading would entail that a statement like » $\phi$  is right if and only if  $\phi$  has property  $F$ « was logically equivalent to the statement » $\phi$  has property  $F$  if and only if  $\phi$  is right«. However, this is contrary to the intended meaning. Explanatory definitions like the first statement are meant to express that  $\phi$  is right *because*  $\phi$  has property  $F$ , i.e., that  $F$  is more fundamental than being right and, thus, explains *why*  $\phi$  is right or is what *makes*  $\phi$  right – and not the other way around. To improve the precision, thus, one could use »if and only if, and if, then because« (or the more common »if and only if, and because«, cf. Timmons [2001]) to explicate the explanatory direction. However, due to verbosity, the shorter version will be used in this document, always implying the more elaborate interpretation for explanatory definitions.

the action is right or wrong, i.e., it is not just the statement of a contingent, extensionally equivalent property:

**Definition 2.2 (Criterion of Rightness)** *A criterion of rightness is a classifying and explanatory statement that for each individual decision situation  $D$  in its domain  $I_T \subseteq I$  specifies a sufficient and necessary condition under which the options of the agent in  $D$  are right and why.*

Here is MOAC's criterion of rightness in this technical, isolated sense:

**Principle 2.5 (MAXIMIZATION)** *For a given decision situation involving an agent  $A$  with a set of options  $\Phi$  and a set of relevant contexts  $C$ : An action  $\phi \in \Phi$  is right for  $A$  if and only if, relative to  $C$ , there is no alternative action  $\phi' \in \Phi$  with better consequences than  $\phi$ .*

Note that this formulation intentionally leaves open which contexts are relevant. As such, it stands for one clearly separated module of MOAC theories. If we combine the objective stance, the first characteristic module of MOAC, with MAXIMIZATION, we get the already known

**Principle 2.2 (MOCoR)** *Let  $D$  be an individual decision situation involving an agent  $A$  with a set of options  $\Phi$  and with actual context  $C$ . An action  $\phi \in \Phi$  is right for  $A$  if and only if, relative to  $C$ , there is no alternative action  $\phi' \in \Phi$  with better consequences than  $\phi$ .*

In this sense, it becomes clear that MOCoR is indeed a composite of the objective stance and MAXIMIZATION as a criterion of rightness. It is, therefore, absolutely correct to say that MOAC theories are precisely those consequentialist theories that embrace MOCoR, and, at the same time, to

say that this determines two out of three consequentialist modules. With this, we can finally turn to the third module.

### 2.3.3 Axiological Background Theories

Finally, we can turn to the axiological module of consequentialist theories. Axiology is the philosophical study of value. Axiological theories are theories about what things have value and how much value they have, whereby »value« refers to value in an intrinsic, non-instrumental sense, sometimes also called »final value« (cf. Schroeder [2021]). Traditionally, consequentialist theories adopt so-called welfarist axiologies, i.e., accounts holding that nothing but well-being matters morally.

In the tradition established by Parfit (Parfit [1984], Heathwood [2020]), we can distinguish three types of welfarist axiologies: Hedonism, Preference Theories, and Objective List Accounts. I do not intend to engage with this debate here since the CHALLENGE is independent of any account of welfare or axiology (in the narrow sense), even though some of the examples I explore might involve the presupposition of a specific axiology. However, they could be rewritten in a way that works with any other plausible axiology – but I do not intend to enter the question of what makes an axiology plausible, either. The important thing is that one does not have to commit to any specific axiology at all when discussing the CHALLENGE. Whichever axiology might be correct (whatever that means), the CHALLENGE arises independently from that question. Thus, I follow a path other consequentialists have taken earlier: I assume that the question of what account of inherent value is settled or, at least, that it can be left to others. In line with tradition, I presume a welfarist account in this book, though.

There is another, *broader* sense of axiologies, according to which they, in addition to the narrow sense, are also concerned with some method(s) of aggregation.<sup>31</sup> After all, consequentialists are, in a sense, more interested in the *value of consequences* than in how well some specific individual is off. They typically need an account of what makes some consequences better than others. In other words, the consequentialist frameworks need methods that lift value from the level of individuals to the level of consequences or even the level of sets of such consequences (if we take, for instance, a subjective stance). Throughout this book, it should be apparent from the context whether I refer to axiologies in the narrow or in the broad sense whenever they are mentioned.

It is convenient to have names for the two parts of the axiologies in a broader sense:

**Grounding Part:** The central question to be answered by the grounding part is: what has how much value (and why)? This part identifies *the value within* consequences. I call this part the grounding part since it grounds the value of (sets of) consequences: a (set of) consequence has a certain value *because* of the value within that consequence. This part corresponds to axiologies in the narrow sense.

**Aggregation Part:** The central question to be answered by the aggregation part is: How good is a particular consequence or set of possible consequences? The aggregation part, thus, is instrumental in illustrating the translation of the value within consequences – as defined by the grounding part – into the overall value of the consequences (or sets

---

<sup>31</sup>There are attempts to develop non-aggregative consequentialist theories (cf. Gustafsson 2021). These approaches are irrelevant in the context of this thesis.

of consequences). Such lifting of value from the individual to (sets of) consequences necessitates at least two forms of aggregation: interpersonal and intertemporal. Here, *interpersonal* aggregation refers to the methodologies for amalgamating values across individuals, while *intertemporal* aggregation denotes the methodologies for consolidating values over time. In general, simple summation is the favored approach for these forms of aggregation, particularly when the scope is confined to finite populations and temporal spans.<sup>32</sup> Furthermore, when it comes to ›interpossible‹ aggregation, which is concerned with elevating the value of consequences to the value of sets of possible consequences, subjective consequentialists often opt to compute the sum of the values of consequences, weighted by the subjective probability assigned to each respective consequence (effectively calculating *expected values*). This approach finds its roots in decision theory (cf. von Neumann and Morgenstern [1947]; Jackson [1991]).

Neither the grounding nor the aggregation part plays a significant role when it comes to *formulating* the CHALLENGE. However, as I am going to argue in the second part of this project, the aggregational part *does* play a crucial part in solving the CHALLENGE. Since this part of the book tries primarily to better understand the CHALLENGE, for now, we only need to be in a position to decide what MOAC recommends in certain situations and to evaluate whether this satisfies MOAC's own expectations. We can, therefore, simply postulate that outcomes of certain actions and combinations of actions have specific values without the need to specify any concrete axiology. What

---

<sup>32</sup>For infinite time horizons, one typically introduces a discount factor such that moments further in the future count less than temporally closer moments.

constitutes the moral quality of outcomes and how it does this does not matter for the ›valuative profile‹ of the cases.

Thus, all that matters is that cases with certain ›valuative profiles‹ exist, but this can be plausibilized pretheoretically, purely based on common sense. Recall FACTORY from above. That it is terribly bad to destroy the river's ecosystem and, in doing so, the livelihood of the village downstream should be beyond doubt and arguably is *ceteris paribus* true with any plausible axiology; the same holds for the assumption that the loss of jobs is awful in a country with underdeveloped social safety infrastructure. Also, it should be obvious that it is better if none of the harmful consequences occur than if they all occur should be clear regardless of our precise notion of what makes one state of affairs better than another. Further, we certainly will agree that a situation where only some of the bad consequences occur is somewhere in between the two extremes. This is all we need to formulate the CHALLENGE – and axiologies that would not be able to account for these judgments would, by all appearances, have to be rejected, anyway.

Thus, the grounding part actually remains irrelevant for this endeavor, and so it makes sense to continue to stay on the level of MOAC, i.e., a *family* of theories that vary in at least this one part of their axiological module (instead of switching to a very specific such theory, like, say, classical Utilitarianism).

Before we put all three modules together and pinpoint the essence of consequentialist theories, it is worth expanding the formal framework developed so far to include axiologies and assessments. This will pay off in the aforementioned second, constructive part of the thesis.

Even though MOCOR, strictly speaking, only requires an axiology that allows to distinguish the consequences with the best qualities from all other

consequences, it is typically assumed that the axiology implies a valuation function, i.e., a function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  that assigns arbitrary outcomes from

$$\mathcal{W} := \bigcup_{D \in \mathbb{I}} \mathcal{O}_D$$

a value from some value space  $\mathcal{V}$ , i.e., the set or domain of values (as induced by some grounding part of an arbitrary axiology). For a set of outcomes  $\mathcal{O}_D \subseteq \mathcal{W}$  of some decision situation  $D$ , let us call  $\text{Val}(\mathcal{O}_D) := \{ \text{Val}(O) \in \mathcal{V} \mid O \in \mathcal{O} \}$  the *valuative profile* (of  $D$ ).<sup>33</sup> In the following, I require the existence of a total order  $\leq_{\mathcal{O}_D}$  over  $\text{Val}(\mathcal{O}_D)$  for arbitrary  $D \in \mathbb{I}$ .<sup>34</sup> Even though, in principle, weaker constraints could be considered and actually will be considered in the context of this project, this assumption of the existence of total orders for arbitrary decision situations reflects quite common, often only implicitly made assumptions.<sup>35</sup> Typically, it is assumed, somewhat oversimplifyingly, that  $\mathcal{V} = \mathbb{R}$  and that the order is just the standard less-equal relation over the reals (with Fehige 1995 being a notable exception). Finally, let  $\text{Val}_C(\phi)$  be agreed upon as an abbreviated notation for  $\text{Val}(\text{Out}_C(\phi))$  (given an individual

<sup>33</sup>For some set  $X' \subseteq X$  and a function  $F : X \rightarrow Y$ ,  $f(X') \subseteq Y$  denotes the *image* of  $f$ , i.e., for the set  $\{ f(x) \in Y \mid x \in X' \}$ . This allows us to introduce some properties of functions that I assume to be known within the context of my project, namely, that a function is called *injective* if and only if for every  $x, x' \in X$ , the fact that  $f(x) = f(x')$  implies  $x = x'$ ; that a function is called *surjective* if and only if  $f(X) = Y$ ; and that a function is called *bijection* (a so-called bijection) if it is a 1-1 mapping, i.e., it is both, injective and subjective.

<sup>34</sup>A *total order*  $\leq$  over a set  $X$  is a binary relation on some set  $X$  which satisfies for all  $x, y, z \in X$ :

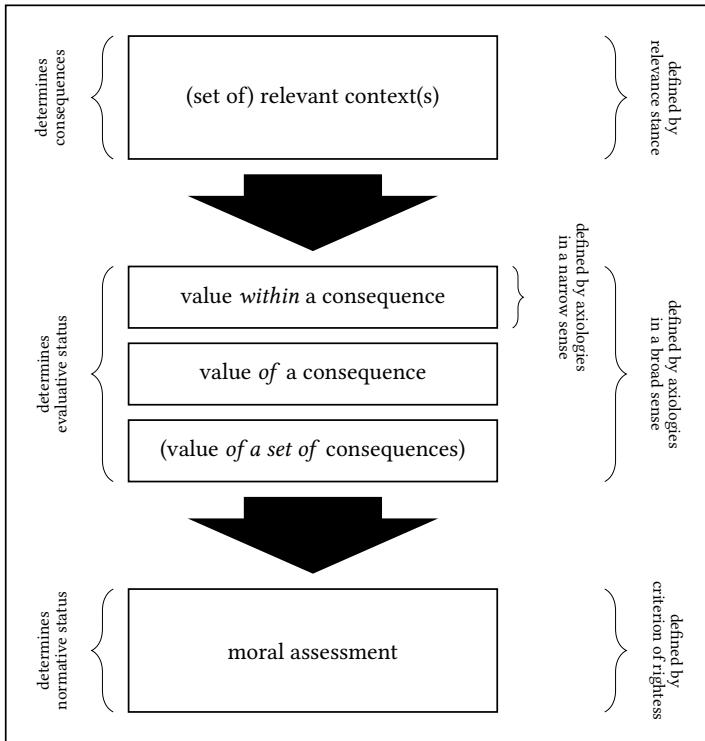
**REFLEXIVITY:**  $x \leq x$

**TRANSITIVITY:** if  $x \leq y$  and  $y \leq z$ , then  $x \leq z$

**ANTISYMMETRY:** if  $x \leq y$  and  $y \leq x$ , then  $x = y$

**TOTALITY:**  $x \leq y$  or  $y \leq x$

<sup>35</sup>MOCOR also works with partial incommensurability as long as there always is a ‘best outcome’, i.e., with partial orders plus some *completeness* property that, for a given  $\mathcal{O}_D \subseteq \mathcal{W}$ , guarantees the existence of a supremum, i.e., a greatest element in  $V(\mathcal{O}_D)$  (with respect to its value). We will later encounter such a conception formulated in terms of being not dominated, see also Harty 2001. For now, we can just ignore that possibility and presume local totality, i.e., the existence of a total order for every  $\mathcal{O}_D \subseteq \mathcal{W}$ .



**Figure 2.5:** How the modules of a consequentialist theory interlink to arrive at moral assessments.

decision situation  $D$  with option space such that  $\phi \in \Phi_D$ , an outcome function Out, and a relevant context  $C$ ).

### 2.3.4 Putting It All Together

Finally, we can put all the modules together and arrive at the *consequentialist assessment pipeline*, consisting of three steps: first, the selection of the relevant context by the relevance stance, yielding, together with the decision situation under consideration, the consequences; second, the evaluation of these consequences in accordance with the axiology; and finally, the moral assessment as a function of the evaluative status of the consequences relative to the relevant context as defined by the criterion of rightness. Figure 2.5 illustrates this pipeline. We can now formulate a formal criterion for qualifying as an objective consequentialist theory:

**Definition 2.3 (Objective Consequentialist Theory (formal))** *T is an objective consequentialist theory if and only if it embraces an axiological sub-theory  $T_{Ax}$  with a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and an objective consequentialist criterion of rightness  $T_{COR}$  such that, for all decision situations  $D \in \mathcal{I}$  and for all  $\phi \in \Phi_D : D, C \vDash_T R\phi$  if and only if  $T_{COR}(\phi)$ .*

*A criterion of rightness  $T_{COR}$  is objective consequentialist if and only if, for all  $D \in \mathcal{I}$  with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  (with  $D$ 's actual context  $C$ )  $T_{COR}$  corresponds to a predicate  $\chi_{T,\text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :*

$$D, C \vDash_T R\phi \text{ if and only if } \chi_{T,\text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))).$$

This is to say that, for a moral theory  $T$  to be objective consequentialist, the rightness predicate implied by  $T$ 's criterion of rightness is extensionally equivalent to a predicate ranging only over the values of options in a decision situation that solely operates on the evaluative profile of  $D$  relative to the actual context  $C$ , i.e., on  $\text{Val}(\mathcal{O}_{D,C})$ .<sup>36</sup> In other words: What makes an action right according to an objective consequentialist theory  $T$  *can only* be the moral quality of the consequences of this action (according to the axiological sub-theory of  $T$ ) in comparison to the moral qualities of the consequences of the

---

<sup>36</sup>In set theory, a *characteristic function* is a function that indicates the membership of elements in a particular set. The characteristic function of a subset  $A$  of a universal set  $U$  – in set theory, the universal set is a set that contains all the elements or objects considered in a particular context or mathematical problem – is defined as follows:

$$\chi_A(x) = \begin{cases} \top & \text{if } x \in A \\ \perp & \text{otherwise (i.e., if } x \notin A) \end{cases}$$

Mathematically, the characteristic function of a set  $A$ , often denoted by » $\chi_A$ «, is defined from the universal set  $U$  to the Boolean domain (or Boolean set)  $\mathbb{B} = \{\perp, \top\}$  (or to the binary set  $\{0, 1\}$ ), so  $\chi_a : U \rightarrow \mathbb{B}$ . So one can freely switch back and forth between the predicate »... is an element of  $X$ « and the characteristic function associated with  $X$ .

Strictly speaking, thus, the predicates or functions characterized here are the characteristic functions of the set of permissible options relative to the moral theory under consideration, the decision situations, and their actual contexts.

other options – and nothing else. This gives us a sharp criterion to distinguish objectively consequentialist theories from other theories precisely.

Obviously, MOAC theories are objective consequentialist theories, not only by name but also according to this definition. In other words, we can state the relevant predicate of MOCOR explicitly as:

$$\chi_{\text{MOCOR},D,C}^{\text{Val}}(\text{Val}(\text{Out}_{D,C}(\phi))) = \begin{cases} \top & \text{if } \forall \psi \in \Phi_D : \text{Val}(\text{Out}_{D,C}(\psi)) \leq \text{Val}(\text{Out}_{D,C}(\phi)), \\ \perp & \text{otherwise} \end{cases}$$

which is logically equivalent to the slightly shorter

$$\chi_{\text{MOCOR},D,C}^{\text{Val}}(\text{Val}(\text{Out}_{D,C}(\phi))) = \begin{cases} \top & \text{if } \phi \in \arg \max_{\phi' \in \Phi_D} \text{Val}(\text{Out}_{D,C}(\phi')), \\ \perp & \text{otherwise} \end{cases}$$

That we can specify such a function for some moral theory  $T$  proves that, in the relevant sense,  $T$  is an objective consequentialist theory. We will see, however, that we can by no means specify such a function for all modifications of MOCOR that have been proposed in reaction to the CHALLENGE. Since this project subscribes to the search for an objective consequentialist solution to the CHALLENGE, Definition 2.3 will function as an important building block of a central criterion of adequacy for this thesis.

Before we turn to the reconstructive task of anchoring the CHALLENGE in its strongest form in the literature – and later the formulation of criteria for assessing moral theories as proposed solutions to the CHALLENGE –, we shortly return to the question of possible *wrongness* predicates, linking back to the discussion above regarding the CONSEQUENTIALIST STANDARD VIEW (cf. Section 2.2 page 30). Although we postpone the discussion of the pros and cons of the following two possible explications of MOAC’s wrongness predicate until later, when they become relevant within the context of

the second part of this project, we take the opportunity, while we are thinking about the rightness predicate, to think also about the concrete form of the wrongness predicate (or, better, predicates).

The first variant,  $W_s$ , is the *easy way*. We simply define wrongness as a shorthand for not being right, i.e.

$$D, C \vDash_T W_s \phi \text{ if and only if } D, C \not\vDash_T R\phi$$

We can translate this back to

$$D, C \vDash_T W_s \phi \text{ if and only if } \chi_{T,D,C}^{\text{Val}}(\text{Out}_C(\phi)) = \perp.$$

I shall call this the *shallow consequentialist wrongness predicate*. We can contrast it with the *deep consequentialist wrongness predicate*  $W_d$ . For this, we define:

$$\bar{\chi}_{\text{MOCOr},D,C}^{\text{Val}}(\text{Val}(\text{Out}_{D,C}(\phi))) = \begin{cases} \top & \text{if } \exists \psi \in \Phi_D : \text{Val}(\text{Out}_{D,C}(\psi)) > \text{Val}(\text{Out}_{D,C}(\phi)), \\ \perp & \text{otherwise} \end{cases}$$

Based on this »anti« version of  $\chi_{\text{MOCOr},D,C}^{\text{Val}}$  we can then define

$$D, C \vDash_T W_d \phi \text{ if and only if } \bar{\chi}_{T,D,C}^{\text{Val}}(\text{Val}(\text{Out}_C(\phi))) = \top.$$

As long as we have a total order over the consequences (via their values) for a decision situation,  $W_s$  and  $W_d$  are obviously extensionally equivalent. Because then:

$$\chi_{T,D,C}^{\text{Val}}(\text{Val}(\text{Out}_C(\phi))) \text{ if and only if } \neg \bar{\chi}_{T,D,C}^{\text{Val}}(\text{Val}(\text{Out}_C(\phi)))$$

because

$$\exists \psi \in \Phi_D : \text{Val}(\text{Out}_{D,C}(\psi)) > \text{Val}(\text{Out}_{D,C}(\phi))$$

if and only if

$$\neg \forall \psi \in \Phi_D : \text{Val}(\text{Out}_{D,C}(\psi)) \leq \text{Val}(\text{Out}_{D,C}(\phi))$$

However, we will later encounter situations where

$$\neg \text{Val}(\text{Out}_{D,C}(\psi)) \leq \text{Val}(\text{Out}_{D,C}(\phi))$$

does *not* entail

$$\text{Val}(\text{Out}_{D,C}(\psi)) > \text{Val}(\text{Out}_{D,C}(\phi))$$

because the options  $\phi$  and  $\psi$  are apparently *incommensurable*. In these situations,  $W_s$  and  $W_d$  will disagree. However, for now, we can leave such formal details and can, finally, turn to the reconstruction of the CHALLENGE.



# **Chapter 3**

## **The CHALLENGE**

This chapter is primarily reconstructive in nature. The principal objective is to anchor my understanding of the CHALLENGE within the context of existing literature. This effort is twofold in its purpose. First, it ascertains that my overall project is grounded in a robust understanding of the subject, thereby minimizing certain dangers, most notably that of attacking a straw man. Second, engaging with the literature is advantageous as it brings to light various conceptual distinctions and allows accumulating classic instances of purported TROUBLEMAKERS. Through the reconstruction of various interpretations of the CHALLENGE and demarcating its scope as addressed in this work, this chapter establishes the foundation that is needed to tackle the CHALLENGE systematically.

This endeavor has proven to be somewhat more laborious than initially anticipated. I believe it has been worth the effort, but I will preface it by explaining why I have chosen my approach and this specific corpus of literature.

### **3.1 On Choosing Giants**

One of the goals of this first part of my thesis is to find relevant accounts of the CHALLENGE anchored in the existing literature and, accordingly, to reconstruct plausible versions of valid arguments. Another goal is to distinguish

the CHALLENGE as addressed in this project from other, related challenges. Finally, I shall cast doubt on the idea that previous approaches have already solved the CHALLENGE satisfactorily for camp MOAC. All of these endeavors are reconstructive in nature and require detailed engagement with a vast body of literature.

Such work is admittedly arduous and exhausting. Nevertheless, it is also crucial – for this particular project and for the progress of science as a whole. If one wants to be a dwarf standing on the shoulders of giants, one should first carefully determine *which* giants' shoulders one actually wants to stand on and how to *best* climb up there. The good thing is that one can try several giants and different paths in different orders to find the best one. Accordingly, I avoid taking the reader up my own nearly decade-long winding path. Instead, I propose a route up, carefully mapped out at the cost of much tears, blood, and sweat. Several years of preliminary work have allowed me to add one or two climbing aids along the way. Hopefully, the result is an enjoyable, if non-trivial, path to new knowledge.

Given that the CHALLENGE has been discussed extensively in the literature for several decades, especially, but not exclusively, in the consequentialist camp, my aspiration comes with a challenge that is typical for reconstructive undertakings: to deal purposefully with a rich and comprehensive corpus of literature going back generations; literature that has not necessarily consistently recognized, let alone referred to each other. At first glance, I am thus faced with a vast, chaotic pile of apparently relevant contributions.

To bring the right kind of order to chaos, one has to define an epistemic goal. Depending on what one wants to achieve, some strategies are more appropriate than others. Since the present project is *not* an undertaking in

the history of ideas, a structural pre-sorting, for example, seems promising in order to prepare a systematic analysis of the CHALLENGE. Accordingly, I will first group the relevant contributions *along the specific variant* of the CHALLENGE they consider.

These two resulting main clusters of work correspond to two strands in the literature that both formulate the CHALLENGE as a genuine (or at least primarily) consequentialist challenge but choose fundamentally different framings for it and, in a way, complement each other in a way that needs to be uncovered. The first cluster frames the CHALLENGE as one of intra-theoretical inconsistency. The second cluster sees it rather as a violation of (consequentialist) intuitions. While the first formulation of the CHALLENGE is based on more presuppositions, it would also be all the more fatal for camp MOAC if successful. Together, the two clusters represent the most relevant contributions in terms of my specific research question.

A different way of dividing these contributions that will occupy us in this chapter provides us with an important distinction that enables a systematic treatment of the CHALLENGE along another dimension. Apart from minor stylistic differences, all authors discuss cases with the same general basic properties – INDIVIDUAL OPTIMALITY and COLLECTIVE SUBOPTIMALITY – that they claim to give rise to the CHALLENGE. I have given these cases the name TROUBLEMAKERS in the introduction. Nevertheless, one can and should distinguish between several different *types* of TROUBLEMAKERS. These types differ with respect to the underlying structure that makes them have the relevant properties. Furthermore, while some contributions focus only on a specific kind of TROUBLEMAKER, others discuss several types. Thus, it seems that to approach the CHALLENGE systemati-

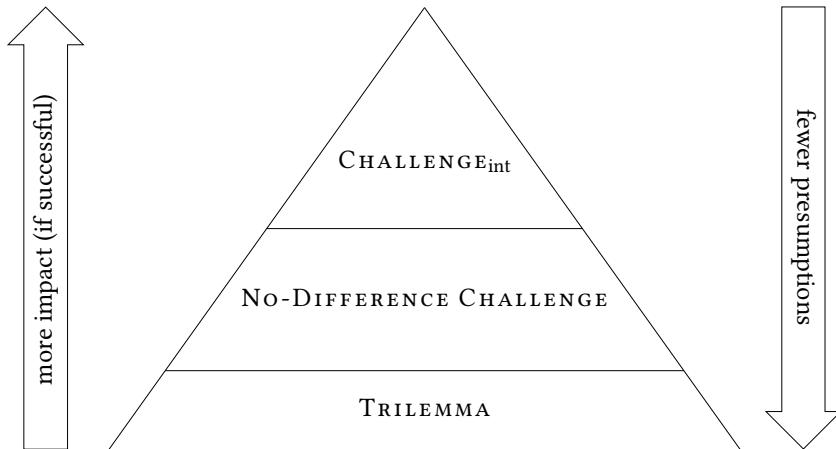
cally, one should have a clear and precise taxonomy of TROUBLEMAKERS. Along the way, I use the opportunity to collect various specimens in this chapter and sketch a taxonomy that will be refined and formalized, at least a little, in the context of the second part of this thesis. Furthermore, along the way, I distinguish the CHALLENGE from two other related challenges.

The number of contributions that concern the CHALLENGE in some way is far too large to treat exhaustively in this thesis. Instead, it makes sense to focus on particularly influential contributions, i.e., pieces that are particularly embedded in and well-connected within the debate, and on contributions that help us draw useful conceptual distinctions – even though this might not do justice to all valuable contributions out there. Therefore, I must decide which works will be examined and to what extent.

All this work will ensure that my project does not burn down a straw man. I hope I have succeeded, but that is for others to judge.

### 3.2 The PYRAMID

This project aims to defend MOAC, and thus a very specific family of moral theories, against the CHALLENGE. In the introduction, I sketched the CHALLENGE in what could be considered its most menacing form for MOAC, which I called CHALLENGE<sub>int</sub>. This particular form, however, is underpinned by quite a number of theoretical presuppositions. As a result, this variant of the CHALLENGE presents a relatively extensive »surface for (relief) attacks« for those in camp MOAC. As previously indicated, however, other variants of the CHALLENGE exist that involve fewer presuppositions but also come with less ›punch‹, i.e., variants that would have less dire consequences for MOAC if successful.



**Figure 3.1:** The three variants that make up the CHALLENGE as the PYRAMID as tackled in this thesis, ordered by their strength and dependence on the preconditions they presuppose.

One question that emerges then is which variant should be subjected to scrutiny. In light of the principle of charity, it is incumbent to grapple with the most promising version of the challenge under consideration. This raises the question of what the CHALLENGE's most formidable version is.

The correct answer, I believe, is actually an evasion of the question. The strongest version of the CHALLENGE is *not one variant* of the CHALLENGE at all, but a *hierarchically ordered sequence of such variants* where one variant lower in the hierarchy backs up those higher up. Only when it is shown that MOAC is not vulnerable to any variant of the CHALLENGE, the CHALLENGE should be considered mastered. I call this hierarchical structure of variants of the CHALLENGE the PYRAMID (cf. Figure 3.1).

The PYRAMID consists first of the CHALLENGE<sub>int</sub>, already outlined in the introduction (and reconstructed with more care in Section 3.5), which has haunted objective consequentialists for decades. In the second row lurks the CHALLENGE in the form of a No-DIFFERENCE CHALLENGE, a formulation that has yet to be reconstructed (this is done in Section 3.4) and is also quite prominent in the literature. Last is a very general formulation of the

CHALLENGE that is not theory-specific, i.e., is meant to threaten a whole set of moral theories – not only consequentialist ones. Due to the chosen representation, which is to be introduced in the following Section [3.3], I call this variant simply the TRILEMMA.

With each additional layer of the PYRAMID, new theoretical assumptions are required to formulate the respective version of the CHALLENGE. On the other hand, the versions represented by the lower levels are based on more assumptions presumed to be ›intuitively convincing‹. The rest of this chapter is devoted to reconstructing the layers of this PYRAMID, starting from the foundation and working our way to the top.

### 3.3 The TRILEMMA

Before delving into the two MOAC-centric variants of the CHALLENGE, which correspond to what I identify as two separate strands within the relevant literature, I will discuss a theory-agnostic variant underpinned by a minimal set of presuppositions. This initial variant is exceptionally well-suited to serve as our point of entry because it predominantly relies on intuitive considerations, allowing us to sidestep the intricate and sometimes nebulous terrain of moral philosophy and, more specifically, consequentialist theorization for the time being.

The variant, which draws inspiration from David Estlund’s formulation (see Estlund [2017] within a non-consequentialist framework,<sup>37</sup> encapsulates

---

<sup>37</sup>Estlund employs the TRILEMMA as a component of a companion in guilt argument to bolster his particular, collective notion of justice. We can ignore the details of his conception of justice, which grapples with a challenge analogous to the one faced here, but it is important to note that Estlund’s TRILEMMA is formulated in terms of moral obligations rather than moral rightness. Consequently, the TRILEMMA discussed here diverges significantly from Estlund’s original version and should not be perceived as semantically equivalent despite being heavily influenced by it.

the fundamental essence of the CHALLENGE through the lens of a trilemma, which, in the interest of brevity, I will refer to as simply the TRILEMMA.

To formulate it we first need a case with the ›right‹ structure. Here is a specific case,<sup>38</sup> borrowed from Felix Pinkert (2015, pp. 973–975), that will serve as the standard example throughout most of this book:

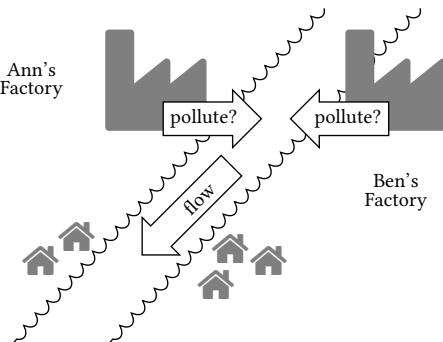


Figure 3.2: The Two Factories case.

**Case 3.1 (Two Factories)** *Ann and Ben each own a factory near the same river. Both can produce either cleanly, or cheaply and thereby pollute the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if a factory closes, this will cause significant unemployment and social hardship. If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream. However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of them can coordinate with the other, nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.*

<sup>38</sup>Estlund uses a structurally identical example involving two doctors, Dr. Stitch and Dr. Patch (cf. Estlund [2017] p. 53–55). However, in my opinion, Estlund’s example is less suitable for presenting this very general variant of the CHALLENGE because it tends to summon certain intuitions about special obligations apparently resulting from the fact that both agents are medical professionals. These intuitions distract from the actual point. Both Pinkert and Estlund’s examples will be introduced and discussed later on. They all belong to a class of cases that is sometimes called »High-Low Cases«, a name that Christopher Woodard attributes at least partly to Michael Bacharach (cf. Christopher Woodard [2003] p. 2017).

TWO FACTORIES involves several agents who are in a certain kind of dependency, which obviously deserves to be investigated in more detail in the course of this chapter. For now, we can note that, to all appearances, this dependency concerns the outcomes (and not the decision-making of the agents which is explicitly *independent*), and it prevents the decision situations from being considered in isolation in a trivial way, i.e., we cannot ›just simply‹ decompose the situation into two individual decision situations, one for Ann and one for Ben. It, therefore, seems appropriate to understand TWO FACTORIES as a description of a *collective* rather than ›just‹ two individual decision situations. The relevant form of dependency is shown mainly by the fact that, whatever the final outcomes of this situation might be, it does not depend on a single action by one of the agents involved, but depends on the specific *combination* of actions that is carried out. Here is a tentative definition of collective decision situations:

**Definition 3.1 (Collective Decision Situation (tentative))**

A collective decision situation *is a situation in which multiple agents are each presented with multiple options, and within a given context, each combination of actions has an associated consequence.*

It will become apparent in the course of this chapter that this definition does not quite do justice to the actual complexity of collective decision situations. However, a more detailed, precise, and even formal elaboration of collective decision situations can be reserved for the second part of this project. For now, Definition 3.1 is sufficient, although it glosses over some issues. Most importantly, we will just have to implicitly assume in the following – much in line with the debate we explore in this chapter – that there is a way to reason about what is the right thing to do for agents in such collective situations.

tions. During the course of this chapter, the common way of consequentialist reasoning in collective situation situations will become more and more clear.

Back to the CHALLENGE. The CHALLENGE in its trilemmatic variant arises as soon as we assume that Ann and Ben both pollute, which apparently constitutes a troublesome combination in the sense defined in Definition 1.1. Recall:

**Definition 1.1 (TROUBLEMAKERS (tentative))** *A collective decision situation is a TROUBLEMAKER if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

**COLLECTIVE SUBOPTIMALITY** *together they would produce a morally suboptimal outcome and*

**INDIVIDUAL OPTIMALITY** *none of them could make a difference for the morally better by unilaterally acting differently.*

The TRILEMMA now consists of the following three statements, all of which appear to be *prima facie* true given the instantiation of that troublesome combination. For example, let us assume that given that both Ann and Ben actually opt for polluting:

- (H<sub>1</sub>) Something wrong happens if the ecosystem of the river and the livelihood of the adjacent village are destroyed.
- (H<sub>2</sub>) If something wrong *happens*, then because someone *did* wrong.
- (H<sub>3</sub>) No one did wrong.

Here are *prima facie* good reasons to believe that all three statements are true. First, given that both Ann and Ben pollute, a lot of harm is done: the river's ecosystem is destroyed, and the livelihood of the nearby village is

eroded. All this is not the result of some calamity of nature or the consequence of a chain of unfortunate circumstances, but a result of *actions*. At the same time, a much better outcome was available through another course of action. If Ann and Ben had both produced cleanly, they would have brought about a situation with all of the benefits of the current one but none of the harm. This clearly would have been better. Therefore,  $H_1$  seems intuitively convincing: The situation that is brought about if both Ann and Ben pollute is a morally unbearable result of actions; something has gone wrong morally.

Second,  $H_2$  seems convincing on conceptual grounds alone. If something is wrong and not merely lamentable, it is because it is a consequence of *action* (see also Principle 2.1 in Section 2.2). Moreover, of course, of at least one wrong action – which must then have been performed by someone, as there can be *no action without agent*. As David Estlund (Estlund 2017) put it: »If something is morally wrong, then there was an obligation on some agent to act or omit other than as they did.« Even if one should find these conceptual considerations questionable,<sup>39</sup> we can simply state the logical consequence of  $H_1$  and  $H_2$  directly. Together, they imply that, in the present case, someone did wrong.<sup>40</sup> Instead of inferring this from  $H_1$  and  $H_2$ , we can simply reduce the two to

$(H_{1,2})$  Some agent did wrong.

In a sense, this is what Frank Jackson does (Jackson 1987, p.100) when he – for a structurally identical case – judges that it »is evident that something wrong

---

<sup>39</sup>It may justifiably be argued that in the case of Estlund's formulation in terms of obligations, the matter is even clearer than in the case of »mere« moral wrongdoing. But this again depends on so many theoretical determinations (cf. Section 2.2) that we need not worry about it here and can put this concern aside for the sake of argument.

<sup>40</sup>Be aware that  $[p \text{ because } q]$  implies not only that  $p$  but also that  $q$  (at least in general – I will actually later argue that this apparent truism should be taken with a grain of salt).

happens[...]; but more than that is evident: something wrong is done. « If we reduce  $H_1$  and  $H_2$  in this way, the TRILEMMA collapses into a very simple dilemma.

Finally, given what Ben actually does, it holds that no matter what Ann would have done, the river would have been polluted. This makes it hard to explain why Ann's act of polluting is an act of wrongdoing, and *vice versa*. The only effect of not polluting would be additional harm through unemployment, worsening the situation even further. In other words, the following conditionals are warranted (against the background of TWO FACTORIES plus the fact that both polluted the environment):

- (4) If Ann had produced cleanly, nothing would have been better, but some things would have been worse because her workers would have lost their jobs.
  
- (5) If Ben had produced cleanly, nothing would have been better, but some things would have been worse because his workers would have lost their jobs.

Thus, they both *did the best they could have done given what the other one did*. None of them could have made a difference for the better by doing otherwise. These considerations certainly strongly support  $H_3$ .

As is the nature of trilemmas, not all of these propositions *can* be true, i.e., this triad of propositions is inconsistent:  $H_3$  denies precisely what follows from  $H_1$  and  $H_2$  (or any other way around). Which one should go? This is the CHALLENGE in pre-theoretic form.

I assume that few would contest  $H_1$ ,<sup>41</sup> while most disagree with  $H_3$ . ( $H_2$  will probably mainly elicit strong opinions from ethicists, who I think will mostly agree.)  $H_3$  seems to be the statement that is indeed somehow off. Michael J. Zimmerman (1996, p. 257) perhaps put the finger on the relevant core intuition when he wrote that it seems as if in such cases there »is a sense in which [...] two wrongs [...] make a right.« The matter would be straightforward if we think about the individual case (recall FACTORY). Polluting would be wrong. But when two such acts come *together*, they seem to oddly »sanctify« each other.

There are undoubtedly plenty of potentially promising starting points for rejecting  $H_3$  on systematic, principle-based grounds, and different moral theories have argued in favor of several of them: Ann and Ben may neglect certain *duties*, say, some duty to not destroy the environment or to do one's share and keep their own hands clean; or they violated certain *rights* of various other parties with their acts of polluting; their behavior may reveal certain *vices*. But which one of these approaches is justified? That depends on what kind of moral theory is correct – and it doesn't look like camp MOAC is in a promising position. Thus, if we want to resolve the TRILEMMA, we need a theoretical framework that allows for deeper analysis. We must leave the cozy realm of pre-theoretical reasoning.

The above considerations apparently in favor of  $H_3$  are essentially based on the observation that no agent could have changed anything for the better

---

<sup>41</sup>It must be pointed out that we are on pre-theoretical ground here. *De facto*, I am well aware of some philosophers who would not accept certain involved formulations. We should not be bothered by this at this point. The only aspect that matters here is that the TRILEMMA corresponds to a (rarely explicated, see especially Estlund 2017) variant of the CHALLENGE that comes up every now and then in the discourse. It does not have a particularly supporting role in this project but forms the final line of defense of the camp CHALLENGE and serves here primarily as a starter, intuition trigger, and pre-theoretic motivational basis.

by acting differently, cf. (4) and (5). Therefore, the problem of refuting  $H_3$  and thus the resolution of the TRILEMMA along the *prima facie* most plausible move, is particularly difficult for theories which we shall call DIFFERENCE-MAKING VIEWS. Such views are characterized by the fact that they accept the DIFFERENCE PRINCIPLE. Here is Frank Jackson's (1987, p. 94) formulation,<sup>42</sup> which should serve us as a starting point:

**Principle 3.1 (DIFFERENCE PRINCIPLE)** *The morality of an action depends on the difference it makes; [i.e.,] it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed.*

Jackson insists that his formulation is attractive for a broad class of moral theories. First, because as the principle »as stated says nothing about how to evaluate the differences, and nothing about what kinds of differences matter morally« (Jackson 1987, p. 94).<sup>43</sup> Second, because it allows that the morality of an action depends *not solely* on the difference it makes, but also on other grounds like, say, intentions or duties. It just states that the moral status of an action (or its »morality«, in Jackson's terms) is *also* dependent on (or a function of) the differences it makes.

Whether this is true or not, it can be said that  $H_3$  in particular should be difficult to reject for DIFFERENCE-MAKING VIEWS. Recall (4) and

---

<sup>42</sup>Jackson's formulation raises some questions. Perhaps the most important for this project is what states of affairs are referred to by »what would be the case were the act not performed.« These and related questions will be neglected for the time being in the interest of the reconstructive endeavor in this chapter. I will address these details in more detail in the second part of this project. For now, we can assume that they are merely a matter of spelling out this relationship specifically, but that there is a reasonable way to spell it out.

<sup>43</sup>It should be obvious that there is a clear connection to the debate about the >consequentializability< of any moral theory, cf. especially Portmore 2007, Portmore 2009, Dreier 2011).

(5) above. If they were indeed both true, it's hard to see how any plausible DIFFERENCE-MAKING VIEW – especially one that is maximizing in the sense that MOAC desires to be – could identify any wrongdoing in TWO FACTORIES (given that the agents actually did pollute). We recall that MOAC theories were characterized by their criterion of rightness, MOCOR:

**Criterion 1.1 (MOCOR (tentative))** *An action is right if and only if there is no alternative action that would actually lead to better consequences.*

Since the consequences of an action are precisely those differences that the action makes in the sense expressed in the DIFFERENCE PRINCIPLE, MOAC theories, of course, are DIFFERENCE-MAKING VIEWS. After all, any objective (act-)consequentialist theory can be characterized by an ›enhanced‹ version of the DIFFERENCE PRINCIPLE:

**Principle 3.2 (Consequentialist's Creed)** *The morality of an action depends solely on the difference it makes; i.e., it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed, and on nothing else.*

Accordingly, the TRILEMMA is particularly hard to solve for MOAC – at least if its supporters really wanted to attack  $H_3$ .

The question then is how heavy the TRILEMMA weighs. Maybe it just shows us something interesting about non-consequentialist intuitions underlying the agreement with  $H_1$  and  $H_2$ . It could be, for example, that consequentialists make the following considerations, which may undermine  $H_2$ : say  $H_1$  merely expresses that there would have been a better possible outcome in the sense that it could have been produced by both agents (cf. COLLECTIVE SUBOPTIMALITY in Definition 1.1). Then we should rather say that

something *needlessly bad* happened. This is perfectly compatible, as TWO FACTORIES shows, with the observation that there are cases where, given what the agents actually do, no individual agent could have produced a better outcome (cf. INDIVIDUAL OPTIMALITY in Definition 1.1). Then, so consequentialists might infer, it does *not at all* follow from the fact that something needlessly bad happened that there must be a wrong action by any agent that explains that badness (or wrongness).

But perhaps MOAC's inability to reject the TRILEMMA shows more. It might show that MOAC theories cannot do justice to basic *consequentialist* intuitions. This leads us to the formulation of the CHALLENGE as the No-DIFFERENCE CHALLENGE, which is consistent with the first, weak strand found in the literature. Alternatively, this inability of MOAC could indicate an even more profound failure in the sense of the CHALLENGE<sub>int</sub> outlined in the introduction, as claimed by the contributions to the second strand. Both traditions of framing the CHALLENGE – the two main variants of the CHALLENGE – are explored below.

### **3.4 The CHALLENGE as No-DIFFERENCE CHALLENGE**

According to the first *weak* strand (which is weak in that it involves fewer assumptions than the CHALLENGE<sub>int</sub> sketched in the introduction), the CHALLENGE arises when certain actions are *intuitively* wrong (right), but the intuitively right (wrong) alternative action, presumably, would make (or would have made) no difference for the better (worse). Looking back to the TRILEMMA and the TWO FACTORIES, the CHALLENGE would then be that it is assumed that  $H_3$  has to be rejected – but at the same time, one cannot do

so because of the apparent individual inefficacy. Call this apparent inability of MOAC to come to *intuitively* adequate assessments in TROUBLEMAKERS the NO-DIFFERENCE CHALLENGE.

In other words, the NO-DIFFERENCE CHALLENGE locates the CHALLENGE in the apparent commitment of consequentialist views to *counterintuitive* moral assessments given the apparent inability of individual agents to make differences in certain cases. Thus, these allegedly violated intuitions are explicitly >consequentialist in spirit<. Here is a quite recent formulation by Holly Lawford-Smith and William Tuckwell (2020, p. 635):

It is an objection to act consequentialist views (but also other difference-making views) that there are classes of actions that don't make a difference and yet we seem to have a strong intuition that those actions should not be performed. Our intuitions suggest that these actions are wrong; the argument from no difference (including insignificant difference) is that they're not wrong because they make no difference. Cases where the actions of many different people add up to cause harm at the level of the collective are prominent examples of where the NO-DIFFERENCE CHALLENGE arises[.]

The NO-DIFFERENCE CHALLENGE constitutes an essential part of the general debate about the CHALLENGE, which alone is reason enough to study it in the context of this project. But even apart from this, engaging with the NO-DIFFERENCE CHALLENGE debate proves to be a fruitful endeavor in several ways. In the following, we collect some relevant observations.

First, the NO-DIFFERENCE CHALLENGE is *broader* than the CHALLENGE. This is partly due to the set of cases that raise the NO-DIFFERENCE CHALLENGE differing from the set of TROUBLEMAKERS. Recall, again, the tentative definition:

**Definition 1.1 (TROUBLEMAKERS (tentative))** *A collective decision situation is a TROUBLEMAKER if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

**COLLECTIVE SUBOPTIMALITY** *together they would produce a morally suboptimal outcome and*

**INDIVIDUAL OPTIMALITY** *none of them could make a difference for the morally better by unilaterally acting differently.*

We can put the characterization from the preceding quotation into a form<sup>44</sup> similar to that definition:

**Definition 3.2 (No-DIFFERENCE CASE (tentative))** *A situation is a No-DIFFERENCE CASE if and only if there is at least one agent that can act in a way such that*

**INTUITIVE WRONGNESS** *it is intuitively morally wrong, but*

**INDIVIDUAL OPTIMALITY** *the agent could not make a difference for the morally better by unilaterally acting differently.*

Thus, as also suggested by the last sentence of the above quotation from Lawford-Smith and Tuckwell, the No-DIFFERENCE CHALLENGE can, in principle, arise in the absence of collective contexts.<sup>45</sup>

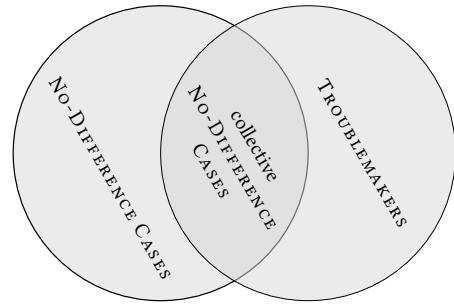
<sup>44</sup>A terminological side note is necessary: I call this class of cases No-DIFFERENCE CASES, simply because it fits with the established name of the challenge they raise, i.e., the No-DIFFERENCE CHALLENGE. But we will see below that this name can quickly be misleading. For example, one might be tempted to think exclusively, or even primarily, of cases such as the climate change scenario described in the introduction, in which, by assumption, each action simply makes no morally relevant difference to some morally catastrophic whole to which individual, insignificant contributions accumulate. Those cases, which we later call CUMULATIVE EFFECTS CASES, are included here but not exclusively addressed. More on this below.

<sup>45</sup>For example, Warren Quinn's puzzle of the self-torturer can be understood not only as one of instrumental rationality but also as a moral problem in the sense of the No-DIFFERENCE CHALLENGE (cf. Quinn 1993). Those who see the self-other asymmetry (cf. Slote 1984) as standing in the way of a moral re-framing of the puzzle may think of a modified version in which one's actions affect another person and not oneself.

Conversely, the definition of TROUBLEMAKERS does not imply that the actions involved are intuitively wrong, but only that the actions that together constitute a troublesome combination lead to *suboptimal* outcomes. Since Definition 1.1 does not mention intuitive wrongness at all but COLLECTIVE SUBOPTIMALITY, we should assume that INTUITIVE WRONGNESS is satisfied in some TROUBLEMAKERS (for at least one involved action), but not necessarily in all.

Thus, the set of NO-DIFFERENCE CASES is distinct from those cases called TROUBLEMAKERS, even if the CHALLENGE is *understood as* a NO-DIFFERENCE CHALLENGE. Thus, that some case is a TROUBLEMAKERS does not entail that it is a NO-DIFFERENCE CASE and vice versa. Nevertheless, the NO-DIFFERENCE CHALLENGE is commonly formulated and discussed in the context of collective decision situations, i.e., it is usually about collective NO-DIFFERENCE CASES (see also Figure 3.3). Whenever I refer to TROUBLEMAKERS in the context of the NO-DIFFERENCE CHALLENGE, I refer to collective NO-DIFFERENCE CASES if not explicitly stated otherwise.

Second, the NO-DIFFERENCE CHALLENGE is broader in another sense as well. It has a plausible variant that also concerns other varieties of act-consequentialist theories, even if we restrict ourselves to collective contexts. In particular, some of the more recent contributions to the debate (Kagan



**Figure 3.3:** Neither are all TROUBLEMAKERS NO-DIFFERENCE CASES, nor the other way around. For this project, only those NO-DIFFERENCE CASES are of interest that are NO-DIFFERENCE CASES due to collective contexts, i.e., that are not only NO-DIFFERENCE CASES but also TROUBLEMAKERS. We can call these cases in the intersection of the two sets »collective NO-DIFFERENCE CASES«. It's stipulated that in this intersection are only cases in which the intuitive moral failure is accompanied by collective suboptimality.

[2011; Hedden 2020] are examples of a collective No-DIFFERENCE CHALLENGE for subjective moral theories. As we will see, this cannot be said for the CHALLENGE<sub>int</sub>. In this respect, a clearer understanding of the difference between the No-DIFFERENCE CHALLENGE and the CHALLENGE helps to frame the present project more clearly and to demarcate it from other collective challenges of consequentialism that fall outside its scope.

Third, the CHALLENGE understood as No-DIFFERENCE CHALLENGE<sup>46</sup> represents a kind of fallback for the CHALLENGE<sub>int</sub> because it works with relatively few presuppositions and, thus, comes with less theoretical baggage than the CHALLENGE<sub>int</sub>. We only need some intuition-based judgments.

Fourth and finally, the discussion of central contributions to the No-DIFFERENCE CHALLENGE also allows us to make some relevant conceptual and theoretical distinctions and to introduce examples that will prove useful in the sequel.

It is a good idea to approach the No-DIFFERENCE CHALLENGE by starting with the contribution that some claim has »introduced [it] to philosophers« (Lawford-Smith and Tuckwell 2020, p. 634), namely with Johnathan Glover's article *It Makes No Difference Whether or Not I Do It* (cf. Glover and Scott-Taggart 1975).<sup>47</sup> Even though this might be a slight overstatement,<sup>48</sup>

---

<sup>46</sup>Hereafter, reference to the No-DIFFERENCE CHALLENGE is always to be understood as referring to the CHALLENGE as or in terms of the No-DIFFERENCE CHALLENGE—should I wish to refer to some instance of the No-DIFFERENCE CHALLENGE that is devoid of collective context, I will make that explicit.

<sup>47</sup>Strictly speaking, this piece has two authors, but it is a bipartite article, and the second part, i.e., the one by Scott-Taggart, is not of interest to this project as he focuses on mere excuses and individual responsibility in collective contexts and not on actual right-doing.

<sup>48</sup>As we will see later, a version of the No-DIFFERENCE CHALLENGE can already be found in a piece by C. D. Broad (1916; see also Section 3.5.1). For the specific version of the No-DIFFERENCE CHALLENGE relevant in the context of this project, the statement by Lawford-Smith and Tuckwell is arguably correct, though. However, No-DIFFERENCE CASES have been discussed, both explicitly and implicitly, earlier in the context of act-consequentialism (cf. Smart and Bernard Williams 1973) discussed in some detail in Sec-

it is fair to say that Glover's contribution remains extremely influential. Accordingly, Glover's account also lends itself as a good starting point for the reconstruction of the CHALLENGE as a NO-DIFFERENCE CHALLENGE.

Glover begins with an observation regarding the existence of what he calls »a family of arguments relating to the insignificance of a single person's act or omission« (Glover and Scott-Taggart [1975], p. 172). He offers two typical statements exemplifying the practice:

- (6) If I don't do it, someone else will.

and

- (7) One person makes no difference.

Glover calls these »attempted justifications«, and the central question of the NO-DIFFERENCE CHALLENGE is whether they actually fail to justify the acts in question. While Glover emphasizes that, in many instances, statements like these are simply false, he agrees that on other occasions, this seems far from obvious. In these cases, the inability to make a difference for the better apparently stands in the way of an intuitively correct moral judgment. In other words, the NO-DIFFERENCE CHALLENGE, like Lawford-Smith and Tuckwell insisted above, indeed presupposes at least the DIFFERENCE PRINCIPLE. (Furthermore, Glover explicitly refers to consequentialism, even though he leaves unclear what exact variant of the theory he has in mind and whether he restricts his considerations to consequentialist theories.)

---

tion [7.3.2.3] and Section [8.3.1]). In particular, the literature on rule-consequentialism is replete with examples of such cases. For instance, R. F. Harrod argued, early in the corresponding debate, that there »are certain acts which, when performed on  $n$  similar occasions, have consequences more than  $n$  times as great as those resulting from one performance« (cf. Harrod [1936], p. 148). Other important examples are given by Gibbard [1965] and Brandt [1959]. We will revisit both later, albeit briefly, as they have inspired J.J.C. Smart and Donald Regan, respectively (cf. Regan [1980], which is discussed in more detail later in this chapter).

Let us return to the two statements and their relation to each other. At first glance, one might wonder whether (7) isn't just referring to a broader class of cases than (6) is, i.e., whether (7) isn't entailed by (6). This is, however, not what Glover had in mind. Instead, properly understood, (6) and (7) cite two distinct reasons for or, respectively, refer to very different structures that give rise to the alleged individual inefficacy. Furthermore, only (6) involves a decidedly collective context. (7), on the other hand, merely refers to the absence of a difference in general, without saying *why* there is no relevant difference made. It gets by without reference to other agents or other actions. Glover, thus, takes the two statements to represent two distinct *types* of collective NO-DIFFERENCE CASES (and of TROUBLEMAKERS).

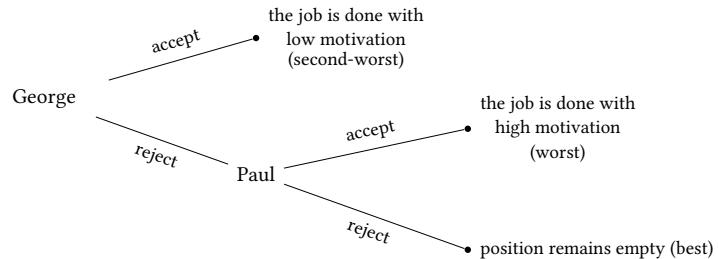
Two concrete examples help us understand the distinction that Glover had in mind. Concerning (6), Glover cites an example from Bernard Williams (cf. Smart 1973, p. 124, modernized a bit by me in the following):

**Case 3.2 (JOB MARKET)** *George, a family man with two children, who has just taken his PhD in chemistry, finds it extremely difficult to get a job. He is not very robust in health, which cuts down the number of jobs he might be able to do satisfactorily. The current situation makes him and his family significantly worse off than if George had a decently paid job. An older chemist, who knows about George's situation, says that he can get George a decently paid job in a certain laboratory, which pursues research into chemical and biological warfare. George says he cannot accept this since he is opposed to chemical and biological warfare. The older man replies that he is not too keen on it himself, come to that, but after all, George's refusal is not going to make the job or the laboratory go away; what is more, he happens to know that if George refuses the job, it will certainly be offered to a contemporary of George's, Paul, who is not inhibited by*

*any such scruples. If Paul were to get the job, he would push along the research with greater zeal than George would. Indeed, it is not merely a concern for George and his family, but (to speak frankly and in confidence) some alarm about this other man's excess of zeal, which has led the older man to offer to use his influence to get George the job.*

If George took the job, he would invest his time and labor into the research and development of chemical and biological weapons, which we might, plausibly and for the sake of argument, understand as an unacceptably bad outcome. However, if George were not to take the job, the much more motivated Paul would take it instead and would put significantly more effort into this harmful endeavor. The best outcome would obviously be that the job remains vacant.

We can visualize the situation by representing JOB MARKET in what is called an *extensive form*, that is, a tree-like graph-based structure like this:



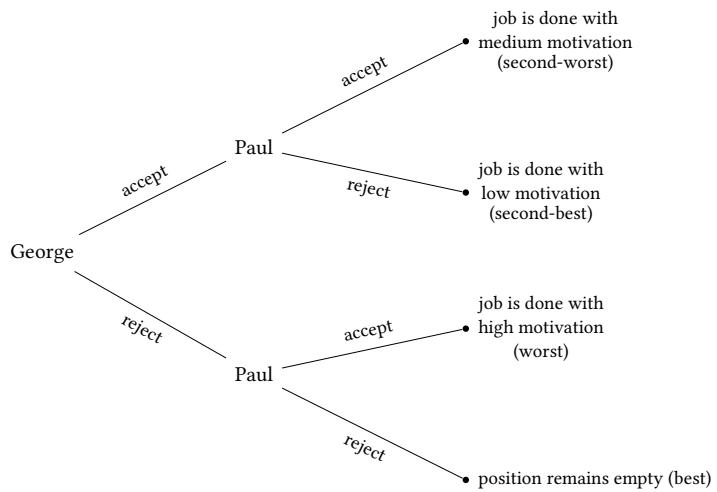
Extensive forms explicitly and visually encode the order of actions: George has to decide first; then, if he rejects the offer, Paul has to make his decision. Let us call such collective decision situations where by description, the order of action is defined (and crucial) SEQUENTIAL CASES.

Strictly speaking, JOB MARKET qualifies as a NO-DIFFERENCE CASE (and as a TROUBLEMAKER) *only* if we are willing to stretch our understanding of combinations of actions such that they are allowed to encompass even hypothetical, counterfactual actions like Paul's. Because then the combina-

tion of George accepting the job offer and Paul counterfactually accepting the job in the purely hypothetical case of George turning it down apparently constitutes a troublesome combination: That George takes the job results in a suboptimal and (let's assume) an intuitively unacceptably bad (and certainly suboptimal) outcome. After all, in principle, a better outcome could have been brought about, namely, if both George and Paul refused the offer and the position remained open. Let's agree that it is intuitively wrong for George to accept such an offer. However, it is not under George's control to bring about the only better outcome. If George did *not* take the job, the much more motivated Paul would be offered it instead – and Paul would take it, with worse results. Therefore, to all appearances, JOB MARKET is a collective NO-DIFFERENCE CASE (and a TROUBLEMAKER), and George can utter (6) truthfully. However, George cannot express (7) truthfully. For although he cannot make a difference for the better, his action *does* make a difference. After all, it would be worse if he didn't take the job because then Paul would step in.

We can, of course, paraphrase JOB MARKET such that we can do without hypothetical actions as components of troublesome combinations. For this, we can imagine that the chemical weapons company actually only needs one employee to pursue its harmful and damnable plans, but they have taken the precaution of advertising two jobs because there is more than enough money to be made from chemical weapons anyway and in times of a shortage of skilled workers, you take every chemist you can get. Assuming that both of them accept, George could still have a moderating effect on Paul's overzealousness for evil; if only Paul accepted, he could do as he pleased, which would have worse consequences; if only George accepted, he would still contribute

something to the bad cause, but much less than they would together. Last but not least, it would be best if the position remained vacant. Because of the good relations with George's older colleague, they offer him the job first. This modified version of JOB MARKET corresponds to the following extensive form:



Other examples with very similar structures are cases of two killers shooting the same person or two perpetrators poisoning a victim and shooting him after poisoning and before the poison takes effect (Parfit [1984]; Jackson [1987]; Jackson [1997]; Zamir [2001]).

All these cases are *minimal* in the sense that they involve two agents, each with two courses of action. Fewer agents and the collective character would be lost; fewer options and there would be no more decisions to make. However, larger cases also fall into the category of SEQUENTIAL CASES. Admittedly, it is hard to get one's head around larger cases, especially cases with numerous agents.<sup>49</sup> The most famous such case is Kagan's chicken example (cf. Kagan [2011], heavily modified below for the sake of simplicity):

<sup>49</sup>With respect to the >size< of such cases, we restrict to the dimension of the number of agents and ignore cases with more numerous options.

**Case 3.3 (CHICKEN COUNTER)** *Shelly is thinking about becoming a vegetarian. He usually buys exactly one chicken a week from his friend Tommi's local deli and doesn't eat any other meat at all. From various casual discussions with Tommi, he knows how the business works: Tommi gets a certain number of chickens delivered from the local poultry farm every week which roughly corresponds to the number of sold chickens the week before. The details are like this: There is a digital counter in his cash register that is set to 0 when a new delivery arrives. Whenever Tommi has sold 10 chickens, his cash register automatically sends a notification to the local poultry farm. This notification causes them to breed (and thereby effectively torture) another 10 chickens. If no new orders come in, they do not breed (and thereby effectively torture) any new chickens.*

Of course, this scenario is completely artificial and in many ways extremely simplistic. But it is enough for us to capture the essential intricacies of real consumer decisions in the form of a TROUBLEMAKER. For this, we need a few more assumptions, though:

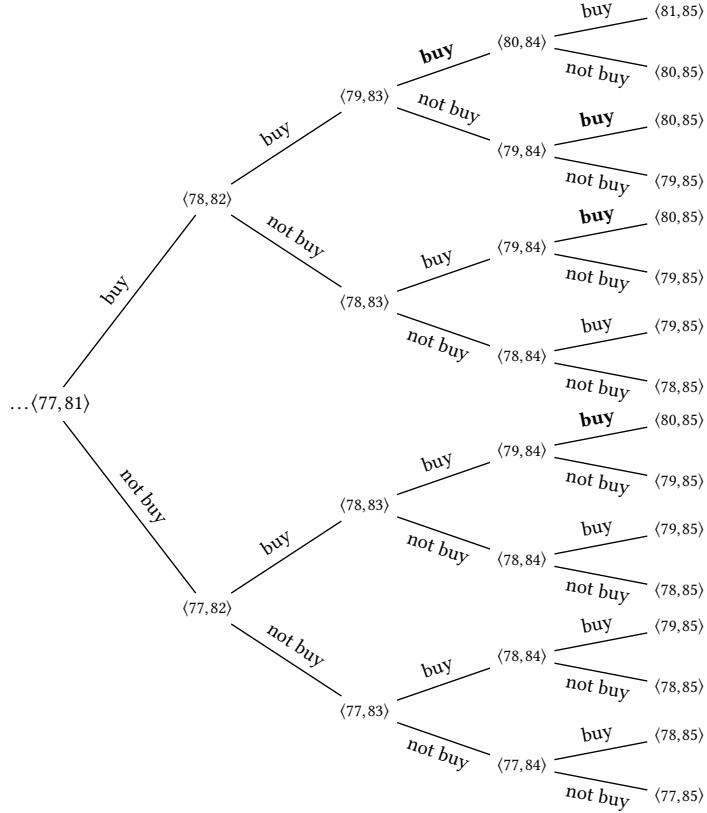
First, we assume the following distribution with respect to the actual decisions made for or against chicken purchases this week:

**Fact 3.1** *85 people considered buying a chicken at Tommi's local deli this week, 81 of which purchased a chicken.*

We also make two assumptions regarding the distribution of moral values:

**Fact 3.2** *Every chicken consumption is accompanied by at least some positive moral value (e.g., pleasure).*

and



**Figure 3.4:** An extract of a possible manifestation of CHICKEN COUNTER for 85 consumer decisions (starting from the 82nd decision). A node labeled  $\langle x, y \rangle$  corresponds to a situation where  $y$  decisions so far have resulted in  $x$  chicken purchases. Buying decisions in bold show actions that triggered the order of 10 more chickens.

**Fact 3.3** *The sum of the individually experienced pleasures of eating chicken is less than the individually experienced agonies of raising chicken.*

Figure 3.4 shows a tiny extract of the corresponding collective decision situations. Note that a node labeled  $\langle x, y \rangle$  corresponds to the situation where  $y$  decisions have led to  $x$  chicken purchases so far. » $\langle x, y \rangle$ « can thus be read as » $x$  of  $y$  visitors have opted for a chicken purchase«. In light of Fact 3.1, we may stipulate<sup>50</sup> that the uppermost path, ending in the state  $\langle 81, 85 \rangle$ , corresponds to the *actual* combination of actions.

<sup>50</sup>To be able to single out *the* path that corresponds to *the* actually performed combination of actions, we actually would need more information, namely which customer bought and which did not. Giving such a list would be a bit lengthy, and I spare us this and further complications. We may simply assume, without loss of generality, that said path corresponds to the actually performed combination of actions.

This is a troublesome combination of actions. First, it is collectively suboptimal: we can quickly identify a significant number of possible combinations of actions that lead to some better result. While the exact number and composition of these combinations depend on the concrete distribution of benefits and harms that is not specified in detail here, we can say for sure that any state  $\langle x, y \rangle$  with  $x \leq 9$  will lead to a better outcome. After all, according to Fact [3.2], the corresponding consequences would involve some pleasures of eating chicken but no additional harm, as no new chickens would be ordered (and thus not raised). Thus, these corresponding consequences have an overall positive value. To the contrary, Fact [3.3] warrants that the actually produced outcome, corresponding to a  $\langle 81, 85 \rangle$  state, is overall negative.

Second, the state  $\langle 81, 85 \rangle$  is individually optimal: For each and every involved agent, indeed, refraining from their chicken purchase would only cost the corresponding amount of pleasure, but not spare any chicken its cruel future.<sup>51</sup> After all, if any individual agent acted otherwise, 80 instead of 81 chickens were consumed, and, thus, the very same number of chickens would have been raised.<sup>52</sup>

CHICKEN COUNTER makes it particularly clear that the SEQUENTIAL CASES discussed so far, i.e., SEQUENTIAL CASES that are TROUBLEMAKERS are specifically involve certain triggers<sup>53</sup> or *thresholds*. These triggers

---

<sup>51</sup>To be precise, forgoing her chicken purchase costs them some pleasure only comparatively, i.e., insofar as it interferes with their chicken consumption (their source of value, by assumption, cf. Fact [3.2]). We can easily make this further complication irrelevant by stipulating that there are no further sources of (dis-)value involved. (Suppose, for example, that some people are feeling bad for choosing *not* to buy chicken. Such affective reactions may sound implausible at first, but I think they can actually explain many a behavior we observe in reality.)

<sup>52</sup>Thus, any combination of actions ending in a state  $\langle x, y \rangle$  with  $x \bmod 10 = 0$  is *not* a troublesome combination of actions.

<sup>53</sup>Shelly Kagan calls them »trigger cases« (cf. Kagan 2011).

are activated by some of the actions that constitute some troublesome combination, respectively; however, other actions *would* have taken the place of each action had it failed to occur. I shall call such cases THRESHOLD CASES.

Figure 3.5 shows the relevant relation of kinds of cases

THRESHOLD CASES are TROUBLEMAKERS qua involving some kind of »adverse causal dependency«, but the underlying nature of these dependencies is importantly different. I will call

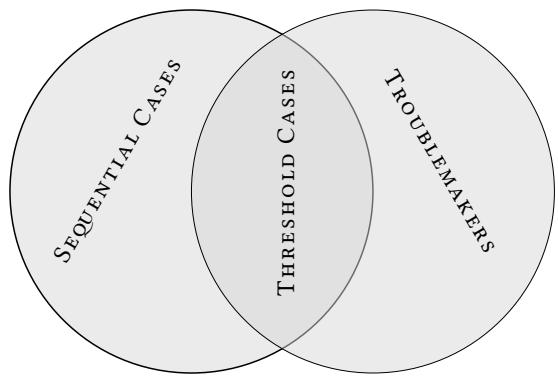


Figure 3.5: THRESHOLD CASES are SEQUENTIAL CASES that are TROUBLEMAKERS.

MARKET *preemption* and structures like in CHICKEN COUNTER (given an unfortunate number of actual purchases) *overdetermination*. Overdetermination and preemption are both concepts in the philosophy of causation, specifically addressing situations where multiple factors appear to be involved in causing a particular effect. Even though there is no single generally accepted definition, here is a rough-and-ready characterization of the two concepts that suffices for the context of this work. Overdetermination occurs when an effect has multiple sufficient causes, any of which would be enough on its own to bring about the effect, and they are all actual (i.e., they all occur). Thus, in overdetermination, multiple actual causes independently suffice to bring about an effect, and they all actually occur. Each cause is, in a sense, redundant; the effect would have happened without it due to the other causes. Preemption occurs when a potential cause is rendered non-actual or irrelevant by another, prior, or more direct cause. In other words, while multiple

factors might have been sufficient to bring about the effect, one specific factor effectively ›takes over‹ and becomes the actual cause, thereby preempting the others. Thus, in preemption, one actual cause effectively nullifies another potential cause, making the latter irrelevant in the actual bringing about of the effect.

This manyfold overdeterminacy can be illustrated particularly vividly using **CHICKEN COUNTER** as an example. given that 81 chicken purchases have been made, the breeding and treatment of chickens are indeed overdetermined by the actions of all the customers collectively. The causal chain that leads from chicken purchases to chicken breeding is activated many times over by the collective actions of these customers. Hence, the overall process of chicken ordering is not sensitive to the actions of any one customer. This indeed creates a situation where individual inefficacy arises from overdetermination – each individual's action is rendered inefficacious by the sufficiency of other individuals' actions for the same result.

George's situation in **JOB MARKET**, however, does not perfectly fit the typical mold of preemption (but the modified version does). George's potential refusal of the job doesn't prevent the job from being done because Paul would take it in his stead. But George taking the job and Paul taking the job wouldn't have the same consequences, i.e., both would have different effects. However, we might stretch the term to capture the underlying structure in **JOB MARKET**. George's refusal is rendered causally inert by Paul's willingness to take the job. After all, his willingness functions as a (genuinely potential) preemptive cause for the continuation of the research, even in a possibly more dangerous direction, regardless of George's decision. This effectively ›bypasses‹ the potential for the best consequence that might

have been actualized if none of them took the job. George's potential to bring about the best consequence is nullified by Paul's readiness to step in, in much the same way that a (classical) preemptive cause nullifies the effect of another potential cause. This highlights an interesting aspect of the case: George's moral stand has a kind of ›fragility‹ akin to the fragility of preempted causes – it doesn't seem to make a difference due to the structure of the situation, much like a preempted cause doesn't get to make a difference due to the structure of its causal network. This brings out the tragic element of George's situation, emphasizing the difficulty of making a meaningful moral difference in a world where others stand ready to nullify our good intentions.

I thus believe that the distinction between overdetermination and pre-emption illuminates the structural differences between cases like *CHICKEN COUNTER* and *JOB MARKET*. Beyond this conceptual acuity, however, this distinction does not play a vital role in this project, as it attempts to solve both types of structures using the same approach. Following others (Jackson 1987; Jackson 1997; Zamir 2001; Spiekermann 2014), I will thus use the term **OVER-DETERMINATION CASES** as a generic label for both types of cases.

Compare that to our running example *TWO FACTORIES*. Recall:

**Case 3.1 (Two Factories)** *Ann and Ben each own a factory near the same river. Both can produce either cleanly, or cheaply and thereby pollute the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if a factory closes, this will cause significant unemployment and social hardship. If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream.*

*However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of them can coordinate with the other; nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.*

If we (plausibly) assume that if both pollute, at least one of the two does something intuitively wrong (and for symmetry reasons, we should then assume that intuitively, both are acting wrongly because the situations of Ann and Ben do not seem to differ significantly), Two FACTORIES is a NO-DIFFERENCE CASE.

However, in contrast to JOB MARKET and CHICKEN COUNTER, Two FACTORIES leave the order in which the agents' actions occur unspecified. I call cases like Two FACTORIES, where such order is not specified, COORDINATION CASES. Two FACTORIES cannot be represented in extensive form.<sup>54</sup> Instead, we can represent it in what is called a *normal form*, i.e., in tabular form:

		Ben	
		pollute	produce cleanly
		pollute	second-worst
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

As we know, Two FACTORIES is a TROUBLEMAKER – and so it has a structure that is quite similar to THRESHOLD CASES: Depending on the

---

<sup>54</sup>If one leaves the rather exotic borderline case of actual simultaneous actions aside, one can represent Two FACTORIES as a set of two →possible← extensive forms, for example. Felix Pinkert has suggested this, cf. Pinkert 2015, p. 975, Fig. 2.

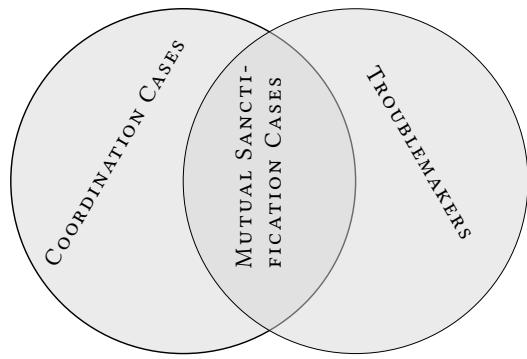
temporal perspective we take, Ann and Ben can truthfully utter different variations of (6), if we assume that both will pollute, do pollute, or already did pollute. In this respect, JOB MARKET, Kagan's CHICKEN COUNTER, and Two FACTORIES all apparently fall into the same category according to Glover's classification – they are all OVER-DETERMINATION CASES.

However, the distinction to

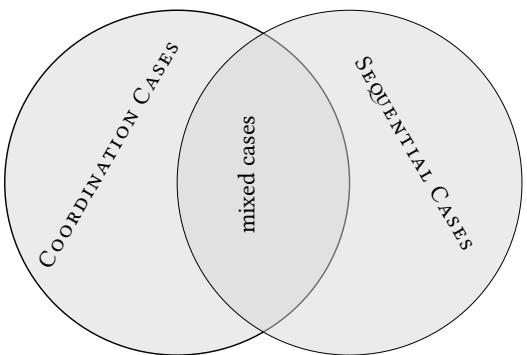
THRESHOLD CASES lies in how the two agents in COORDINATION CASES appear to mutually overdetermine each other's actions. To quote Michael J. Zimmerman again (1996, p. 257), it seems as if in such cases there »is a sense

in which [...] two wrongs [...] make a right.« I will use the term MUTUAL SANCTIFICATION CASES to refer to COORDINATION CASES that are also TROUBLEMAKERS.

Figure 3.6 shows this relationship. In contrast, Figure 3.7 gives an overview over the »space« of collective decision situations, without involving the TROUBLEMAKER question.



**Figure 3.6:** MUTUAL SANCTIFICATION CASES are COORDINATION CASES that are TROUBLEMAKERS.



**Figure 3.7:** Collective decision situations can be COORDINATION CASES or SEQUENTIAL CASES, depending on whether the order of actions is fixed or open. Mixed cases, where the order is fixed for some but not all actions, are possible. I omit them in this project as they are normally ignored in the literature, but that should not worry us regarding generality. Whatever we learn with respect to »pure« cases should be easily transferable to mixed cases.

Two remarks on COORDINATION CASES: first, like THRESHOLD CASES, these cases can involve an arbitrary number of agents. Here is an example with a slightly xenophobic connotation from R. B. Brandt (1959, p. 389) that pitches the CHALLENGE (arguably as NO-DIFFERENCE CHALLENGE) to make a case for Brandt's rule-utilitarianism and that still sounds pretty contemporary:<sup>55</sup>

Suppose that, in wartime England, people are requested, as a measure essential for the war effort, to conserve electricity and gas by having a maximum temperature of 50 degrees F. in their homes. A utilitarian Frenchman living in England at the time, however, argues as follows: »All the good moral British obviously will pay scrupulous attention to conforming with this request. The war effort is sure not to suffer from a shortage of electricity and gas. Now, it will make no difference to the war effort whether I personally use a bit more gas, but it will make a great deal of difference to my comfort. So, since the public welfare will be maximized by my using gas to keep the temperature up to 70 degrees F. in my home, it is my duty to use the gas.«

According to the act-utilitarian theory, this argument is perfectly valid. But we should not take it seriously in fact. Why not? At least part of the reason is that we think that, if a sacrifice has to be made for the public good, all should share in it equally. Imagine the outcry in Britain, if it became known that members of the Cabinet, who knew that electricity and gas were in good supply because of the country's willingness to sacrifice, used this argument to justify using whatever power was necessary to keep their homes comfortable.

---

<sup>55</sup>Not only because it restates once again a variant of the CHALLENGE (I'd say in the form of the NO-DIFFERENCE CHALLENGE), but also because, living in the time of another actual war in Europe, I at least have heard that argument more than once in 2022, when the German Gas reserves were considered insufficient for the winter (cf. specifically Habeck's statement and the surrounding debate mentioned in the Preface, page [v](#)).

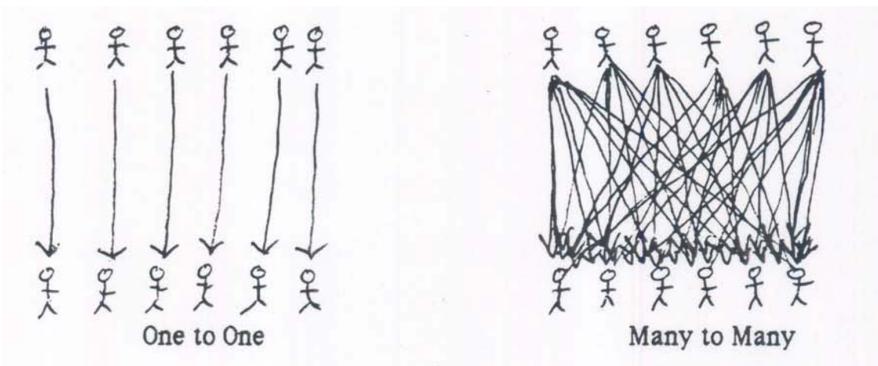
In this case, it is irrelevant and, above all, unspecified who sets their heating to which level and in what order. It is, thus, a COORDINATION CASE.

Second, COORDINATION CASES are *not* to be confused with what is sometimes called »moral prisoner’s dilemmas« (Parfit [1984], chapter 2). While these, like COORDINATION CASES, leave the order of actions unspecified, they presuppose agent-relative obligations or values (or, in Parfit’s terms, »aims«), which, in fact, gives rise to the typical structure of a prisoner’s dilemma, with ›moral payoffs‹ for each agent.

Besides being an example for a *n* person COORDINATION CASE, Brandt’s Frenchman example is also a case where (□) can apparently be uttered truthfully by the involved agents and thus brings us back to Glover’s distinction. Such cases involve actions that, in some sense, make no difference at all or, at most, a negligible difference. In sum, however, many such acts matter a lot.

The most important real-life candidate for such a case has already been mentioned in the introduction: the man-made climate crisis. The individual contribution of a normal citizen aggregated over a whole lifetime is probably measurable and statistically quantifiable. Still, this at most concerns what Julia Nefsky called the »underlying dimension« (cf. Kagan [2011]; Nefsky [2011]): whether some life is lived in an ascetic or wasteful manner probably makes no *moral* difference with regard to climate, even if it might be measurable in some way. Such real-life cases, however, bring into play many empirical questions that we should leave either to climate scientists or even cognitive psychologists (cf. E. N. Dzhafarov and D. D. Dzhafarov [2010b]).

Glover actually offers an own example that allows us to circumvent empirical questions. While Glover formulates this in the form of a long description of two successive alternative resolutions of two structurally identical col-



**Figure 3.8:** Original drawings by Derek Parfit (Parfit 1988 p. 29) illustrating the structure of both BEANS AND BANDITS situations. The bandits are depicted by the stick figures at the top, the villagers by those at the bottom, and arrows indicate acts of stealing (where the sum of effects of all actions directed at one villager in Many to Many equals the effects of one action directed at that villager in One to One).

lective decision situations, it turns out to be purposeful to pull these two sub-cases apart. This way, we get what Derek Parfit once called a »Glover pair« in an unpublished manuscript (Parfit 1988)<sup>56</sup> First, consider (Glover and Scott-Taggart 1975, pp. 173):

**Case 3.4 (BEANS AND BANDITS (One to One))** *Suppose that in a village, there are 100 very hungry, almost starving tribesmen who prepare their lunch on 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one villager's bowl to satisfy their appetite.*

Certainly, no bandit can truthfully utter (7). Each and every bandit's action makes a difference for the worse since it costs a villager their well-deserved and much-needed lunch. Hence not stealing a bowl would, no doubt, make a difference for the better. But now compare this to the second case:

---

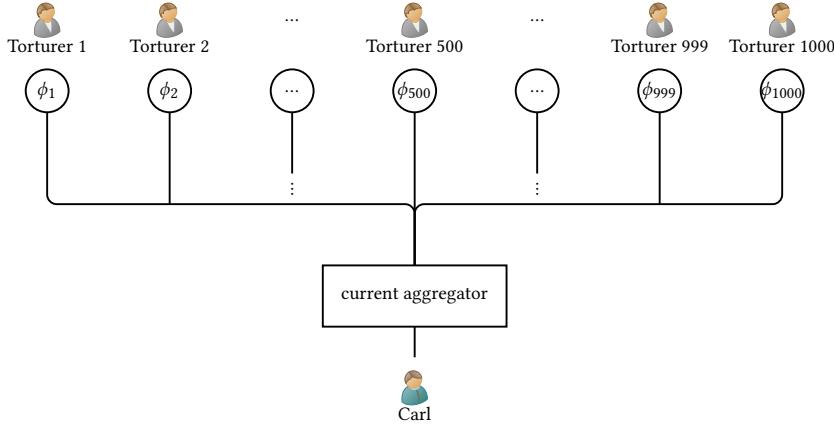
<sup>56</sup>I also simplified the description a bit.

**Case 3.5 (BEANS AND BANDITS (Many to Many))** Suppose that in a village, there are 100 very hungry, almost starving tribesmen who prepare their lunch on 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one bean from each villager's bowl to satisfy their appetite.

Assuming that one bean more or less makes no moral difference to the villagers – no matter how many beans are left – each of the 100 bean thefts committed by each bandit is ›morally inefficacious‹ in the relevant sense. Then, when uttering (7), each bandit would say a truth. Thus, the second version of BEANS AND BANDITS is another candidate for a NO-DIFFERENCE CASE, at least if we assume that stealing the beans is intuitively morally wrong. Thanks to its collective context and general structure, this BEANS AND BANDITS case is also a TROUBLEMAKER. Figure 3.8 shows Parfit's illustration of both situations' structure (cf. Parfit 1988).

Other well-known examples of this class of TROUBLEMAKERS have been introduced by Derek Parfit, and especially his HARMLESS TORTURERS example and DROPS OF WATER still enjoy some popularity. Let us have a look at a modified variant of Kagan's revised version of the HARMLESS TORTURERS (cf. Kagan 2011, p. 116, see also Figure 3.9).

**Case 3.6 (HARMLESS TORTURERS)** Carl is wired to a torture machine with a thousand identical switches. When none of the switches are flipped, no current runs through the machine, so Carl is in no pain. If all a thousand switches are flipped, then a sizable current runs through the machine, and Carl is in tremendous pain (but no permanent damage is done to his body). But the flipping



**Figure 3.9:** The setup of the HARMLESS TORTURERS case: 1000 torturers can each flip their switches  $s_1$  to  $s_{1000}$ ; the number of switches flipped determines the strength of the shock Carl receives.

*of any given switch increases the current only by a tiny amount (well below the perceptually discriminable threshold for pain) so that Carl simply cannot tell whether one switch more or less has been flipped – regardless of how many other switches have already been flipped. Finally, imagine that a thousand different people each control a single switch and must decide whether to flip it or not. None of them cares about Carl or feels any remorse, but each of them enjoys flipping switches.*

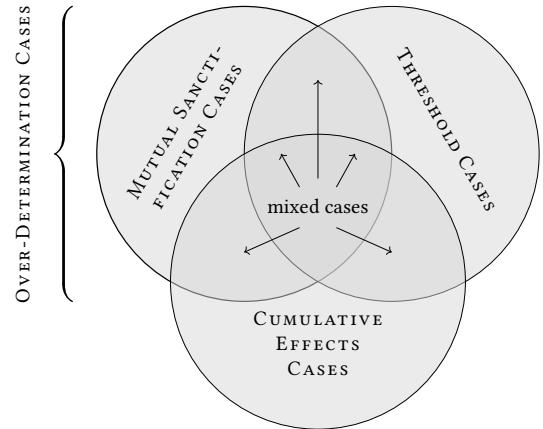
Of course, this is a **NO-DIFFERENCE CASE** (and also a **TROUBLEMAKER**). For if a sufficient number of torturers flip their switches, the resulting pain Carl experiences will outweigh the sum of the slight pleasures of the torturers in flipping their switches. It is evident that in this class of cases, it is relevant that, in some way, a set of individually negligible effects at an underlying level somehow add up to a total effect that outweighs the sum of all benefits arising from the individual actions. Therefore, this class of **TROUBLEMAKERS** cases shall be called **CUMULATIVE EFFECTS CASES** in the following. **CUMULATIVE EFFECTS CASES** are **no OVER-DETERMINATION CASES**.

Figure 3.10 shows the relationship between the different cases.

We can now summarize what we have learned about different kinds

of TROUBLEMAKERS. First, OVER-DETERMINATION CASES are those cases in which INDIVIDUAL OPTIMALITY is due to some form of overdetermination (or preemption) which is indicated by the agents' ability to truthfully utter (6). Second, THRESHOLD CASES are OVER-DETERMINATION CASES in which the order of actions is fixed, i.e., SEQUENTIAL CASES that are TROUBLEMAKERS. Third, MUTUAL SANCTIFICATION CASES are OVER-DETERMINATION CASES in which the order of actions is *not* predetermined, i.e., COORDINATION CASES that are TROUBLEMAKERS. Of course, mixed forms are conceivable, i.e., situations in which the order is fixed for some but not for all. Finally, there are CUMULATIVE EFFECTS CASES and, again, mixed cases with all other kinds of TROUBLEMAKERS.

This rough-and-ready taxonomy of TROUBLEMAKERS proves useful in this project, primarily because the underlying different INDIVIDUAL OPTIMALITY true-making structures by which these types are characterized arguably call for different approaches to the CHALLENGE. This idea goes back at least to Derek Parfit (1984, chapter 3), but continues to enjoy great popularity today (see, for example, Kagan 2011). However, my taxonomy is, in its entirety



**Figure 3.10:** The set of all TROUBLEMAKERS as the union of the three different types of TROUBLEMAKERS. OVER-DETERMINATION CASES are the union of COORDINATION CASES and THRESHOLD CASES. All mixed cases, i.e., all cases in any intersections, are omitted for systematic reasons, both in the literature and in this project. The idea is that, on the one hand, one can master the CHALLENGE more easily for >pure TROUBLEMAKERS< and, once one has working approaches for all these pure type cases, one can also construct complex solutions for mixed cases.

(cf. Figure 3.3, Figure 3.5 and Figure 3.6), more nuanced.

For now, however, we can leave behind the taxonomy and return again to the No-DIFFERENCE CHALLENGE in order to wrap up this section. In a nutshell, the idea of the CHALLENGE as a No-DIFFERENCE CHALLENGE is this: there are collective decision situations where some action is, intuitively speaking, morally wrong. However, because of individual inefficacy, i.e., the inability of the individuals to make a difference for the better by acting otherwise, plausible difference-making views apparently fail to come to the intuitively correct assessment.<sup>57</sup> This version of the CHALLENGE as a No-DIFFERENCE CHALLENGE arguably finds an appropriate representation in the following argument:

**Argument:** The No-DIFFERENCE CHALLENGE ARGUMENT (tentative)

*P<sub>ENDCs</sub>:* There are No-DIFFERENCE CASES: collective decision situations in which there is at least one agent that can act in ways such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

---

<sup>57</sup>Note that I restricted the investigation here to the *negative* formulation of the No-DIFFERENCE CHALLENGE. Alternatively, we could formulate it positively. There are apparently cases where some action does not make a difference *for the better*, and yet we seem to have a strong intuition that those actions are *right*. This positive formulation, to my knowledge, is seldom discussed in academic circles. However, it often surfaces in everyday moral discourse. Consider, for instance, individuals who adopt veganism or consciously reduce behaviors tied to high carbon dioxide emissions. They believe these choices are morally right, even if they might not see an immediate tangible impact. One could reframe the No-DIFFERENCE CHALLENGE to highlight the potential misjudgment of these intuitive stances. Indeed, academic discussions might exhibit a bias, emphasizing the risks of false positives (mistakenly deeming intuitively wrong actions as right) over the potential pitfalls of false negatives (overlooking actions that are intuitively right but are assessed as wrong).

$P_{\text{MOCOR}}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg\text{intu}}$ : If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then  $T$  is counterintuitive.

---

$C_{\neg\text{intu}}$ : MOAC is counterintuitive.

This argument is apparently deductively valid,<sup>58</sup> so we directly turn to the premises. In light of the example cases presented,  $P_{\exists\text{NDCs}}$  seems reasonably backed, at least if we accept, for the sake of argument, the intuitive assessments involved. Thus, the soundness of the argument hinges on whether  $P_{\text{MOCOR}}$  and  $P_{\neg\text{intu}}$  are true. Given MOCOR, it seems indeed hard to deny that  $P_{\text{MOCOR}}$  is true, but we investigate this in detail in the next subsection (and, honestly, in the second part in even more detail) when introducing the CHALLENGE<sub>int</sub> and, thus, the ARGUMENT.  $P_{\neg\text{intu}}$  seems a lot like a conceptual truth, even though one can question whether *one* clash with intuitions should suffice for labeling a whole theory counterintuitive. However, since the NO-DIFFERENCE CHALLENGE makes a rather systematic point and diagnoses such a clash for a broad class of cases, we might accept it nevertheless. In this respect, the NO-DIFFERENCE CHALLENGE ARGUMENT seems relatively convincing – as long as one shares the central intuitions, anyway.

---

<sup>58</sup> Assuming that the class of NO-DIFFERENCE CASES qualifies as a significant class of situations, which seems more than plausible.

At the same time, its appeal to intuitive assessments is also the weakness of the No-Difference Challenge. Since it attacks the intuitive extensional adequacy of a theory, it is, on the one hand, quite broad in its application. It threatens the convictions of any adherent of a theory  $T$  that (in possibly slightly modified No-Difference Cases) come to the relevant assessments, as long as he shares the central intuitions. As we shall see, it can, therefore, also be brought to bear against subjective varieties of act-consequentialism. On the other hand, one can escape the Challenge as No-Difference Challenge if one simply does not share the central intuitions. And even if one does share them, one might be willing to »bite the bullet«: maybe we are learning here, drawing an important general lesson for our endeavor of theory-building. What theory comes without any intuitively questionable judgments?

In other words, if we are to formulate a stronger version of the Challenge that puts MOAC in more serious trouble, we could do so as soon as we could provide reasons why these seemingly intuitive assessments cannot simply be written off by MOAC's advocates. Suppose it can be shown that MOAC is even *committed* to these assessments. In that case, we have an existential challenge because then the diagnosed inadequacy of MOAC would not be that of an intuitive extensional inadequacy but that of a theoretical inconsistency – which is clearly an unbearable burden for any theory. As outlined in the introduction, there is indeed such a theory-internal framing of the Challenge.

### 3.5 The CHALLENGE as CHALLENGE<sub>int</sub>

A much stronger formulation of the CHALLENGE describes it as an intra-theoretical conflict. I have already sketched this formulation in the introduction – in this section, I will dig a little deeper.

#### 3.5.1 Starting from a Non-Starter: The Principle of False Universalization

One of the earliest modern analytic engagements with TROUBLEMAKER-like cases is found in a 1916 essay by C. D. Broad (1916). Like Glover, Broad grapples with a particular structure of argumentation that often appears in this context, viz.

(8) What if everyone did that?

A brief examination of Broad's considerations proves useful for this project in two ways. First, (8) stands for a practice of justification often put forward in the context of collective action. As we will see, Broad argues convincingly that this is not a suitable practice from a consequentialist point of view, therefore, we too can put it to rest. Second, Broad's discussion of this justification practice introduces a version of the principle MH in the passing and is thus ideally suited for the transition to a stronger framing of the CHALLENGE as CHALLENGE<sub>int</sub>, which heavily relies on PMH. Broad made these points almost 50 years before Glover prompted the debate surrounding the NO-DIFFERENCE CHALLENGE. It is, hence, worthwhile to turn briefly to Broad's reflections.

Contemplation akin to (8) typically ends with the condemnation of the action under consideration because, if everyone were doing the action under

consideration in such circumstances, the consequences would be unacceptable. C.D. Broad (1916, p. 377) pondered on that practice by discussing the specific principle instantiated in such considerations:

A man proposes to himself a certain course of action and debates whether it be right or wrong. At a certain stage he will say to himself, or, if he be discussing the matter with a friend, his friend will say: Suppose *everybody* did what you propose to do. The consequences of this hypothesis will then be considered, and, if they be found to be bad, the man will generally consider that this fact tends to prove that his proposed action is wrong.

Broad called that principle the *principle of false universalization*. It is a principle of *false* universalization as we »are asked to believe that the rightness or wrongness of many of our actions depends on the probable consequences, not of what we judge to be true, but of what we know to be false« (Broad 1916)<sup>59</sup> The principle invites us to consider, given an actual individual decision situation, a hypothetical collective decision situation where all the (hypothetical) agents perform the (hypothetical) action corresponding to the (actual) action that the (actual) agent is considering. Then, the (hypothetical) outcome of everyone – or at least a sufficient number of people to produce some relevant effect – performing ›the action‹ under consideration is evaluated. If the result is judged to be unacceptable (or »bad« in Broad's own terms), then the original action under consideration is deemed wrong. Here is what I take to be a fair formulation of that principle:

---

<sup>59</sup>Broad certainly is correct when diagnosing a certain similarity between the principle of false universalization and Kant's categorical imperative (especially in the so-called law of nature formulation). But there can be no doubt either that there are important differences. Thus all that follows is not to be read as an argument against Kant's theory. Also, note that there are many points of connection to universalizability and impartiality requirements involved in many schools and traditions of moral theorizing.

**Principle 3.3 (FALSE UNIVERSALIZATION (negative))** *Let  $\phi$  be an option that would not make a difference for the worse.  $\phi$  is morally wrong if a sufficient number of type-identical actions would together produce morally unacceptable consequences.*

Broad considers several examples and several possible ways to apply the principle. Here is one particularly interesting passage (Broad [1916], pp. 383):

I walk through a field and pluck an ear of corn. Is this right, wrong or indifferent? If I now say: Suppose a million people walked through and each plucked an ear, the results would be very bad[...].[I]t seems to me that the argument from the damage done by a million ears being plucked to that done by the plucking of one is most precarious. The consequences that have to be considered cannot be the mere separation of the ears from the stalk; this, like all physical events, is in itself morally indifferent. We obviously have to go further and consider the effects on the state of mind of the owner of the field and of others. Now it seems perfectly possible that no one's state of mind is in the least better or worse for the plucking of one ear and yet that it may be very much the worse for the plucking of a million. There is absolutely no logical reason against this and it seems to me to be true. The most probable account of the matter is that the plucking of a certain finite number  $n$  (varying of course with the circumstances) is absolutely indifferent, while the plucking of any greater number leads to consequences which get worse as the number gets greater.

There is obviously a connection between the cases Broad targets here and TROUBLEMAKERS, even though his cases *are no* TROUBLEMAKERS. In Broad's cases, a single individual act makes no relevant difference, but a multiplicity of *hypothetical* acts would have unacceptable consequences. This, then, is considered a basis for condemning the actual, individual act. In

TROUBLEMAKERS, by contrast, a multiplicity of *actual* acts taken together have unacceptable consequences, while the effects of the single acts seem negligible *because*, given the other actual acts, none of them seems to make any significant difference.

What would help with TROUBLEMAKERS, to arrive at the apparently desired assessments, is thus a principle rather like this one:

**Principle 3.4 (TOTUM PRO PARTE)** *Let  $\phi$  be an option that would not make a difference for the worse.  $\phi$  is morally wrong if it is part of a combination of actions that produce morally unacceptable consequences.*

TOTUM PRO PARTE would allow us to assess the individual questions in TROUBLEMAKERS as wrong even though, by definition, no alternative would have made a difference for the better. Of course, people have come up with this approach to the CHALLENGE, and we will briefly discuss and reject this strategy (as proposed by Frank Jackson Jackson [1987]) and its drawbacks in Chapter 4.

FALSE UNIVERSALIZATION has its own problems. In contrast to TOTUM PRO PARTE, it links the outcome of certain *hypothetical* combinations of actions to the moral status of an *actual* option. Especially from an objective consequentialist perspective, it is difficult to see how *merely hypothetical consequences* should carry ›moral weight‹, the principle cannot (or at least not readily) be made fruitful for MOAC. Accordingly, Broad agrees that the principle of false universalization fails because it fails to establish such a robust link. Adopting a rather consequentialist perspective, he even concludes »that both on practical and ethical grounds it is most unlikely that you can ever safely argue from the goodness or badness of the effect of a number of precisely similar acts to the rightness or wrongness of a single act of the

class» (Broad 1916, pp. 383). (Note that if this were true, not only FALSE UNIVERSALIZATION but also TOTUM PRO PARTE would be a non-starter.)

Broad's considerations, however, come with a caveat that brings us back to PMH. He wants to make an exception for an *inverted* variant of FALSE UNIVERSALIZATION that might be used to establish not the condemnation of an action but its *rightness*. Not the bad effects of a large number of hypothetical actions are used as a basis for such assessment, but the *positive* effects of a multitude of hypothetical actions of the same type are taken into account (or, respectively, the negative effects of a continuation of the status quo). In other words, according to that inverted principle, we infer from the morally *welcome consequences* that a multitude of certain *hypothetical actions* would have to the rightness of each such *actual option*. Consider

**Principle 3.5 (FALSE UNIVERSALIZATION (positive))** *Let  $\phi$  be an option that would not make a difference for the better.  $\phi$  is morally right if a sufficient number of type-identical actions would produce morally acceptable (i.e., optimal) consequences.*

In the wild, we find the application of the positive variant at least as often as that of the negative, especially in the context of large-scale challenges such as climate change. Here the principle is often applied not only in the private sphere but in the public discourse as well. Where the negative version is used to tell us to give up our beloved steaks, the positive version is used to reinforce our consumption of oat milk. The two variants, apart from their lack of theoretical persuasiveness (at least from a consequentialist perspective), complement each other excellently.<sup>60</sup>

---

<sup>60</sup>There is certainly a connection to what I called the positive variant of the NO-DIFFERENCE CHALLENGE, cf. Footnote 57.

For some reason, Broad believes this positive variant of the principle holds (*ibid.*, p. 391):

Is there then no valid use for the principle of false universalisation in ethics? I think there is at least one, though it is a very modest one. It can be used to refute a certain kind of mistaken judgment about the rightness of a suggested act. Suppose that certain acts are very unpleasant to everyone and entail very real sacrifices from which everyone shrinks. Suppose further that the performance of such acts by a certain number of persons is essential to the attainment of a considerable good or the avoidance of a considerable evil. If now a man says: I will not act thus *because* I dislike the sacrifice then it is open to us to point out to him that, if this be his sole ground, it is just as valid a ground for all other people, since by hypotheses they all dislike the sacrifice. If then he is right in refusing to do the act, all other people will also be right in refusing on the same ground. But the result will be that a great good will be lost or a great evil suffered.

It is very interesting for the context of this thesis to understand why Broad believes that he and the consequentialists *need* this principle. After all, it seems quite questionable to accept the relevance of hypothetical combinations in one case but not the other. Thankfully, Broad lets us know why he *wants* the positive variant to be true (*ibid.*, pp. 392):

*Now it cannot be the case that the result of a number of right actions can be a state of affairs which can be foreseen to be worse than if people had acted differently* [emphasis added]. Hence we can conclude that these actions could not all be right. But if his ground for supposing that his action was right were valid all these actions would be right.

This means that Broad proposes here to raise the positive principle of false universalization because he thinks that otherwise another principle he

believes to be immutably true would be violated: That it cannot be the case that some combination of exclusively right actions leads to predictably suboptimal outcomes. This, of course, is a variant of PMH from the introduction.

Before I make that explicit, let me emphasize that I do not think that Broad can have it both ways. He claimed (and I think rightly so) that the negative variant of FALSE UNIVERSALIZATION is invalid because we are not allowed to cite the unacceptable *hypothetical* consequences of *hypothetical* actions in support of the condemnation of *actual* individual inefficacious options, since there is a missing link between the hypothetical and the actual. But for the very same reasons, then, we are not allowed to cite the foreseeable *hypothetical* betterness of consequences of *hypothetical* actions in support of the rightness of *actual* individual options that come with certain moral costs.

I am not here to defend or attack Broad's reasoning, though. For my project, the important point is this: Broad considers it beyond doubt that »it cannot be the case that the result of a number of right actions can be a state of affairs which can be foreseen to be worse than if people had acted differently.« For defending *that* idea, he is willing to accept the principle of false universalization for certain cases. That idea, arguably, can be stated as

**Property 3.1 (Broad's Property)** *If all agents do right, then, necessarily, they do not foreseeably produce morally suboptimal consequences together.*

I call Broad's Property a *property* (and not, say, a principle) because it is best understood as a property of moral theories. This property may (or may not) accrue to some moral theories and, of course, involves an implicit universal quantification overall. That a moral theory has this property means that, for all decision situations (within its domain) and under all relevant contexts, this

theory makes such recommendations that following these recommendations, i.e., doing what is right according to the theory in these situations and relative to these contexts, can not lead to foreseeableably suboptimal outcomes.

In other words, Broad believes that, if all agents act rightly, then they are guaranteed to produce together either morally optimal outcomes or morally suboptimal ones, but then this happens in an unforeseeable manner. Broad's formulation, thus, has an epistemic and, thus, subjective twist: The phrase »can be foreseen« immediately makes one wonder for whom it should be foreseeable. From Broad's example, it is clear that he has in mind the agents' perspectives. However, we might want to use EPISTEMIC LIMES to translate Broad's property to an objective stance. Recall:

**Principle 2.4 (EPISTEMIC LIMES)** *Let  $T_O$  be an adequate objective (i.e., non-epistemic) moral theory and let  $T_S$  be an adequate subjective (i.e., epistemic) moral theory. For a given decision situation involving an agent A with a set of options  $\Phi$  and a set of relevant contexts C: If A knows all relevant facts and has no incorrect relevant beliefs, then*

$$T_O, D, C \vDash R\phi \text{ if and only if } T_S, D, C \vDash R\phi.$$

Although EPISTEMIC LIMES must be taken with a grain of salt – as there might be things that one cannot know at the time of acting like what other agents are doing –, we might for a moment assume that we can apply it to drop the subjective aspect which gives us an objectivist's version of Broad's Property. This version then boils down to the collective reading of CONGRUENCE from the introduction:

**Principle 1.1 (COLLECTIVELY MAXIMIZING (tentative))**

*If all agents act rightly, then they are guaranteed to produce the morally best outcome they could bring about together.*

Thus, we have indeed re-arrived at what I earlier called, following Feldman (Feldman 1980), the PRINCIPLE OF MORAL HARMONY (PMH). When Broad writes that »it cannot be the case« that MH is violated, he implicitly states that he took MH as a necessary property of every plausible moral theory. In other words, Broad would have apparently accepted:

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

At this point, let me draw three lessons from Broad: first, consequentialists must not hope that FALSE UNIVERSALIZATION will help to master the CHALLENGE; second, that one of the fundamental principles underlying the CHALLENGE is much older than the discussion of the CHALLENGE itself; third, that this principle is not unique to genuinely consequentialist moral thought (as Broad was no consequentialist, cf Gustavsson 2021).

The following will be devoted to the explicit reconstruction of the ARGUMENT and thus of the foundation of the CHALLENGE as CHALLENGE<sub>int</sub>, its strongest variant.

### 3.5.2 Regan's Impossibility Result

Donald Regan was not one afraid of bold claims. Here is what he sets forth as the goal of his book *Utilitarianism and Co-operation* (Regan 1980, p. vii): »In this essay I shall first analyze and then dissolve a contradiction which the

existing literature suggests is inherent in utilitarian theory and which, if it were genuinely indissoluble, would weigh heavily against the acceptability of any form of utilitarianism.« As my present project is based on the existence of that contradiction, it's important both that it can be resolved and that it has not been resolved successfully by Regan in 1980. So we should follow Regan's thoughts in quite some level of detail. He sketches his plans further:

[T]here are two distinct and equally compelling particular intuitions subsumed under the general utilitarian intuition that moral agents should be required to maximize good consequences. According to one of these particular intuitions, each individual agent should be required to act in such a way that the consequences of his own behaviour are the best possible in the circumstances confronting him as an individual. According to the other of these particular intuitions, any group of agents should be required to act in such a way that the consequences of their collective behaviour are the best possible in the circumstances confronting the group as a whole. [...]

In the course of this essay, I shall show both that the problem I have just indicated is a real problem, and that there is a solution. In the first half of the essay, I shall demonstrate that the two particular intuitions I have identified [...] can not be reconciled by any moral theory of the general sort which utilitarian theorists have proposed up to now.

It should be obvious that the two allegedly ›irreconcilable intuitions‹ correspond to MOCOR and COLLECTIVELY MAXIMIZING, while the ›general utilitarian intuition‹ refers to what I called

**View 1.1 (CONGRUENCE)** *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

Regan claims that he demonstrates that »the two particular intuitions [...] can not be reconciled«. It is this ›impossibility result‹ that corresponds to the CHALLENGE<sub>int</sub> as sketched in the introduction and that shall be reconstructed, partially in combination with similar arguments by other authors, in more detail in this section.

It should be noted that Regan does not target merely MOAC theories but a broader set of victims. To understand the scope of his effort, some of Regan's ›technical lingo‹ proves helpful. Most importantly, he distinguishes ›exclusive act-orientation‹ theories – corresponding to what he called ›the general sort‹ in the above quote – from all other approaches to moral theories. While Regan sees himself unable »to produce a definition of exclusive act-orientation which is both precise and completely general« (Regan [1980], p. 109), he gives a characterization of the property for the context of COORDINATION CASES (*ibid.*, p. 113): »whether an agent satisfies a traditional consequentialist theory depends, ordinarily, on what he does from a list of acts [...] and not on what he tries to do or how he decides what to do. This is the feature of traditional theories I refer to by saying they are ‘exclusively act-oriented’«. Much more recently, Douglas Portmore offered a more general definition: »a theory is exclusively act-orientated if and only if it requires only that agents perform and refrain from performing certain voluntary acts« (Portmore [2018], p. 14). Whatever is the best way to put the distinction, it should be clear that MOAC indeed is such a theory by the very structure of MOCOR alone. But even if we can somehow save MOAC from Regan's attack, it is important to remember that it may continue to affect every other exclusively act-oriented theory. However, it will turn out that it is not at all trivial to get right to the heart of the intuition behind this property.

While Regan's work did not attract a particularly wide audience, Derek Parfit made it the basis of one very influential chapter of his *opus magnum Reasons and Persons* (Parfit 1984).<sup>61</sup> Accordingly, Regan's work echoes via Parfit's and informs current works on the CHALLENGE like Pinkert (2015); Portmore (2018) among others. To my knowledge, Regan's discussion of the CHALLENGE is the most thorough up to this day and thus deserves a particularly close look in the context of this work. After all, as mentioned earlier, my work can be seen in some respects as a reissue of Regan's project.<sup>62</sup> Yet, as similar as our goals are – to analyze the CHALLENGE and, more specifically, the CHALLENGE<sub>int</sub> in detail and then to solve it – our results are different.

However, this thoroughness comes at a price. Although Regan limits his investigations to only one subvariant of TROUBLEMAKER, namely COORDINATION CASES, his work draws heavily on self-defined notions, abbreviations and various case distinctions. For at least some of the distinctions, a rough and ready understanding is necessary in order to understand his main argument. Because we can learn much about the strong strand of the CHALLENGE by examining Regan's reasoning carefully, I take some time to unfold his argument of which we have seen the preliminary version already in the introduction:

---

<sup>61</sup>Another distinct yet related strand of literature where Regan's work resonates lies within a specific area of game theory. Michael Bacharach's work on the so-called *Hi-Lo Cases* (Bacharach 1999; Bacharach 2006) is notably influenced by Regan. (Both Regan and Bacharach were significantly inspired by Thomas Schelling's analysis of conflicts and coordination problems (Schelling 1980; see also Regan 1980, pp. 133, 191, 198, 260, 265).) Thus, Regan's influence continues to be evident in contemporary discussions within the branch of game theory concerned with cooperation and ›team reasoning‹, as demonstrated in recent works (Petersson 2017; Gold and Colman 2020).

<sup>62</sup>At the same time, Regan's project is heavily inspired by the already mentioned article by Allan F. Gibbard, cf. Gibbard (1965).

**Argument:** The ARGUMENT (tentative)

*P<sub>3T</sub>*: There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome,

but none of them could make a difference for the better by unilaterally acting differently.

*P<sub>MOCOr</sub>*: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

*P<sub>MH</sub>*: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

*C<sub>¬Adeq</sub>*: MOAC is not an adequate moral theory.

In what follows, I reconstruct this argument step by step, always starting with Regan but also looking clearly beyond Regan as a primary source. This will be done in three steps, each corresponding to one of the premises (but in a different order than in the argument, reflecting Regan's own story-telling).

### 3.5.2.1 Step 1: WHIFF AND POOF and P<sub>ET</sub>

Regan puts a very simple and artificial example<sup>63</sup> at the center of his book (Regan 1980, p. 18):

**Case 3.7 (WHIFF AND POOF)** Suppose that there are only two agents in the moral universe, called Whiff and Poof. Each has a button in front of him which he can push or not. If both Whiff and Poof push their buttons, the consequences will be such that the overall state of the world has a value of ten units. If neither Whiff nor Poof pushes his button, the consequences will be such that the overall state of the world has a value of 6 units. Finally, if one and only one of the pair pushes his button (and it does not matter who pushes and who does not), the consequences will be such that the overall state of the world has a value of 0 (zero) units. Neither agent, we assume, is in a position to influence the other's choice.

WHIFF AND POOF is a COORDINATION CASE that can be represented in the following normal form:

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

<sup>63</sup>As Regan attributes correctly, this example is basically identical to of Alan F. Gibbard: »The situation will be as follows. Jones and Smith sit in their isolation booths with red push-buttons. If at 10:00 a.m. both are holding down their push-buttons, they receive cake and ice cream, which is intrinsically good. If only one of them is holding his push-button down, however, they both receive electric shocks, which is intrinsically bad. If neither of them is holding his button down, nothing happens«, cf. Gibbard 1965, p.215, where it used to make roughly the same point Regan makes, but with the goal to shed light on the relation between rightness in act- and in rule-consequentialist terms.

It should be obvious that WHIFF AND POOF is structurally equivalent to TWO FACTORIES in the sense that they have the same outcome profile. Consequently, WHIFF AND POOF is a MUTUAL SANCTIFICATION CASE. For now, a rather intuitive understanding of the notion of identical outcome profiles will suffice: For every combination of actions in WHIFF AND POOF there is a combination of actions in TWO FACTORIES such that the corresponding mapping is stable with respect to the order of the values of the corresponding outcomes. Ann and Ben both producing cleanly corresponds to Whiff and Poof both pushing, etc.<sup>64</sup> Thus, all we learn from Regan's argument for WHIFF AND POOF with respect to MOAC can be straightforwardly transferred to TWO FACTORIES – and, more generally, to COORDINATION CASES. After all, according to MOCOR, the order of the values of outcomes of actions is all that matters with respect to the moral assessment of actions (as we have defined in Definition 2.3 at the end of the preliminary chapter).

Both WHIFF AND POOF and TWO FACTORIES seem to offer coherent, consistent and sufficiently complete descriptions of collective decision situations. Thus, in light of the lightweight criterion for the existence of cases in the relevant sense (cf. Section 2.1, page 25) they apparently warrant the first premise of the ARGUMENT, viz.

---

<sup>64</sup>It is not worth the work to spell out this definition formally here. Later, we will have to do so for the sake of an important argument.

(P<sub>ET</sub>) There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

The next component of Regan's argument is considerably more interesting.

### 3.5.2.2 Step 2: PROPCOP and P<sub>MH</sub>

The second theoretical piece in Regan's puzzle is (Regan [1980], pp. 4):

**Property 3.2 (PROPCOP (Regan))** *If all agents satisfy T in all choice situations, then the class of all agents produce by their acts taken together the best consequences that they can possibly produce by any pattern of behavior.*

According to Regan, PROPCOP<sup>65</sup> PROPCOP is one of the »two particular intuitions [that] can not be reconciled« (see above). For Regan, PROPCOP is a (at least historically) widely accepted criterion of adequacy for consequentialist theories (and arguably for other theories than exclusively act-oriented theories, the already mentioned property I discuss in the next subsection in more detail).

We already stumbled upon a very similar idea when considering Broad's passage above, and one stumble upon similar ideas again and again in centuries of literature on moral philosophy, especially in the context of consequentialist thinkers. The following is a short excursion into the corresponding

---

<sup>65</sup> »Prop« stands for »Property« and »COP«, which stands for a certain theory that is meant to have that ›cooperative‹ feature articulated in PROPCOP (Regan [1980], p.85):

**Principle 3.6 (COP)** *An act is right if and only if it is prescribed (for the agent whose act is in question) by that universal prescription for action, the universal satisfaction of which would produce the best possible consequences.*

history of ideas in order to ensure that the notion I offered above does justice to the general idea.

We start with Fred Feldman, who, in the very same year Regan published his book, published an article (Feldman 1980) on what he called

**Principle 3.7 (PRINCIPLE OF MORAL HARMONY (Feldman))** *When all the members of a social group do what they morally ought to do, the group as a whole does benefit more than it would have from the performance of any worse alternative set of actions.*

The similarity between Regan's PROPCOP and Feldman's PRINCIPLE OF MORAL HARMONY is certainly striking. While Fred Feldman does not refer to Regan, he does refer to a variety of other proponents of very similar expectations either of morality in general or specifically of consequentialist theories. For instance,<sup>66</sup> take Berkeley's view in his *Passive Obedience* (Berkeley 1712, p. 239). Operating under the assumption that judgments of rightness are grounded in a set of moral rules (or »precepts« in Berkeley's terms) that he called ›the law of nature‹ which express God's judgments, Berkeley claims that »the law of nature is a system of such rules or precepts as that, if they be all of them, at all times, in all places, and by all men observed, they will necessarily promote the well-being of mankind, so far as it is attainable by human actions.« Furthermore, we already encountered Jeremy Bentham's take on the issue in the introduction. He took his *Principle of Utility*, »which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question«, to be »capable of being consistently pur-

---

<sup>66</sup>The following list is heavily inspired by Feldman's discussion of the principle in said article, cf. Feldman 1980

sued« and believes that »it is but tautology to say that the more consistently it is pursued, the better it must ever be for humankind« (cf. Bentham [1780]). In more modern terms, we can find a similar idea in Stephen Toulmin's work (Toulmin [1953], p.137), who once wrote that »we can provisionally define [the ›function‹ of ethics] as being 'to correlate our feelings and behaviour in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible'.« Feldman collected several other examples of philosophers that embraced *some* version of PMH, some of which I have not yet mentioned, most importantly Kurt Baier (Baier [1958]), Hector-Neri Castañeda (Castañeda [1974]), and J.L. Mackie (Mackie [1977]).

Even though Feldman himself tried to argue against this principle, the idea remains alive. For instance, when Michael J. Zimmerman discusses the concept of moral obligation, he writes<sup>67</sup> in his chapter on cooperation (cf. Zimmerman [1996], chapter 9)):

Thus, given [an account of moral obligation that demands everyone to do the best one can, call it *T*], *neither* [Whiff] *nor* [Poof] does wrong in [not pushing], even though [...] the best that *both* can do is *not* done. [...] This implication of [*T*] should surely trouble anybody inclined to the view that one ought to do the best one can, inasmuch as it demonstrates that there is a sense in which universal satisfaction of [*T*] is compatible with the best that can be done *not* being done.

Similarly, Douglas Portmore (2018, p.13) formulated the idea as an explicit criterion of the correctness of moral theories, drawing explicitly from Regan's ADAPTABILITY (to be discussed in the next section):

---

<sup>67</sup>In the following adapted in order to match WHIFF AND POOF instead of his specific case.

**Criterion 3.1 (MH (Portmore))** *A moral theory T is correct if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could together bring about.*

Finally, we can use the opportunity to return to Felix Pinkert – the inventor of Two FACTORIES – who writes (Pinkert [2015] p. 975):<sup>68</sup>

Act Consequentialism judges that both Ann and Ben act rightly by polluting, even though they together could easily have brought about much better outcomes by both producing cleanly. So if Consequentialists only have Act Consequentialism to morally apprise [sic!] a situation, then they let Ann and Ben off the hook for together producing collectively suboptimal outcomes. This is at odds with the following claim:

**Principle 3.8 (ON-THE-HOOK)** *In any collection of agents who together gratuitously fail to bring about collectively optimal outcomes, there must be some relevant morally objectionable facts about some of the agents.*

Pinkert has particular strong opinions with respect to ON-THE-HOOK (*ibid.*, pp.976):

[...] ON-THE-HOOK is a widely shared assumption in the philosophical discussion of Consequentialism and no-difference cases. [...] ON-THE-HOOK should not be understood as a specifically Consequentialist position. Instead, it should be understood as the contraposition of a second-order claim about morality in general and, hence, as a desideratum for any moral principle. According to this claim, the relation between morality and overall value is such that if everyone

---

<sup>68</sup>Note that I leave Pinkert's principle untouched in both naming and wording, but pull it out of the citation so that it can be numbered consecutively in line with the other definitions and principles.

always satisfied all requirements posed on them by morality, the world would be as good as it can be (as far as agents' influence is concerned). [...]

Thus understood, ON-THE-HOOK has considerable intuitive appeal, and a moral principle that can accommodate this intuition is, other things equal, strongly preferable to a moral principle that cannot accommodate it.

This second-order claim with »intuitive appeal« Pinkert evokes here is certainly *a collective reading* of CONGRUENCE, i.e., a version of PMH. In contrast to MH, Pinkert's ON-THE-HOOK allows compromises concerning the possible fulfillment of this second-order claim by adding the formulation »gratuitously«. Pinkert writes (*ibid.*, pp. 976): »By calling a failure to bring about optimal outcomes 'gratuitous,' I mean that the failure cannot be explained by mitigating circumstances due to which we could not expect a given group to collectively act optimally. Typically, such circumstances consist in non-culpable misinformation or lack of information. Since Ann and Ben know all relevant facts, I assume that their failure to bring about optimal outcomes is gratuitous.« Arguably, Pinkert is assuming here a principle in the spirit of EPISTEMIC LINES (cf. Principle 2.4, page 43), claiming that this would allow subjective accounts to embrace ON-THE-HOOK, too<sup>69</sup> – something that is implausible for MH to assume.

It is worth devoting a brief digression to the qualification Pinkert has in mind and why an unqualified formulation of MH is not appealing for subjective accounts because it will turn out later that there is no way around a qualified formulation of MH. For this purpose, we briefly leave the playing field of MOAC theories and enter that of MAXIMIZING SUBJECTIVE ACT-CONSEQUENTIALISM (MSAC).

---

<sup>69</sup>Pinkert's ON-THE-HOOK is thus in direct line with Broad's Property, cf. Section 3.5.1.

MSAC is a family of subjective consequentialist moral theories, i.e., all of its members come with a subjective stance (as introduced in Section 2.3, View 2.2) that takes into account the epistemic situation of agents.

As becomes clear by a glimpse at their criterion of rightness, the MAXIMIZING SUBJECTIVE CRITERION OF RIGHTNESS (short: MSCoR), they are *prospective* theories:

**Principle 3.9 (MSCoR (prototypical))** *It is right to perform a certain action if and only if there is no alternative with expectedly better consequences.*

Prospective theories like MSAC theories do certainly not *generally* embrace MH. It is helpful to have an example at hand that shows why this is the case. Consider the following, provided by Frank Jackson (Jackson 1991).<sup>70</sup>

**Case 3.8 (THE DRUG)** *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of the other two drugs, either B or C, will completely cure the skin condition; the other though will kill John, and there is no way that she can tell which of the two is the perfect cure and which is the killer drug.*

---

<sup>70</sup>Parfit made a similar point in an unpublished piece (with a case sometimes called the MINERS, cf. Parfit 1988) some years earlier. While it is plausible to assume that Parfit's case motivated Jackson's, a very similar case can be found in Donald Regan's book. However, Regan's example is hidden in a footnote at the very end of a book on the CHALLENGE (Regan 1980, pp. 264–265, footnote 1 of Chapter 11) where he also already concludes that »[in] case the reader is worried by the fact that my practical suggestions sometimes lead agents to abandon the attempt to produce the best consequences theoretically possible, I note that the same is true of a sensible practical approach to the application of any consequentialist theory.« I add all this just to emphasize that considerations of this kind are quite close to the CHALLENGE: both Parfit 1988 and even more so Regan 1980 are important sources with respect to the CHALLENGE and will enter the stage again and again.

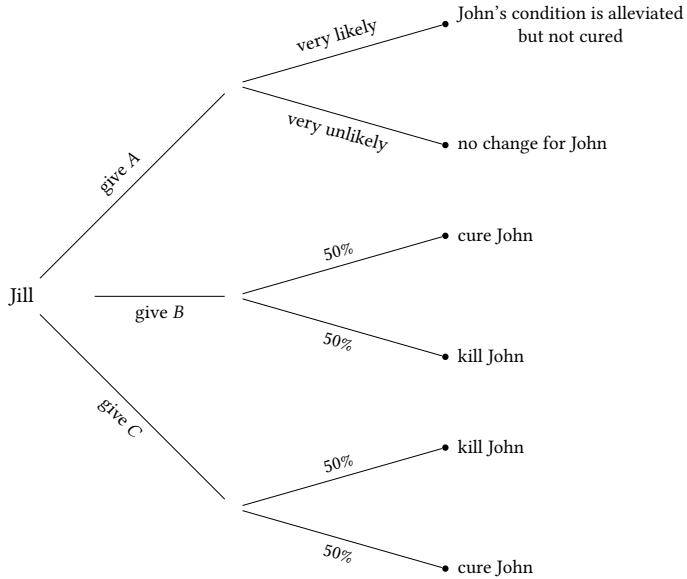


Figure 3.11: The extensive form of THE DRUG.

We can depict this case in the extensive form depicted in Figure 3.11.

Even without concrete calculations in terms of expected value, it seems safe to estimate that giving *A* has the best expected consequences. Consequently, according to MSAC, it is right for Jill to give *A*. At the same time, Jill does *know* that giving *A* does not have the best consequences. After all, only *B* or *C* may cure John. As soon as we accept that the subjective, epistemic situatedness of an agent plays a role with regard to the question of what is morally right for them, this judgment also seems to be the intuitively correct one. Taking a gamble and freely choosing between *B* and *C*, however, seems morally condemnable. Even if Jill were lucky enough to choose the curative drug, it would seem appropriate to reproach her morally. To play a fifty-fifty cure vs. death lottery in the presence of a minor albeit non-trivial harm seems to be negligent, especially in the presence of an option that will very likely alleviate the condition without the risk of causing additional harm.<sup>71</sup>

<sup>71</sup>A SUBJECTIVE VIEW like Railton's (Railton 1984), according to which, very roughly, one ought to aim for the best outcome, apparently would recommend either to give *B* or *C*, but definitely not to give *A* (as this excludes the best for sure). As this seems rather implausible, it

Many lessons can be learned from examples like THE DRUG. But the one essential takeaway with respect to the CHALLENGE is that subjective theories are typically happy to embrace the fact that there can be situations where right-doing even *excludes the possibility* of bringing about the morally best outcome. Even though THE DRUG teaches us this lesson with respect to individual decision situations, there isn't an obvious reason why this should not be true for collective decision situations as well. Once a theory does consciously dismiss CONGRUENCE it arguably might, *ceteris paribus*, dismiss the two readings of it as well.

Certainly, Pinkert's formulation of ON-THE-HOOK seems rather convincing at first glance since it excludes precisely those circumstances that make guaranteed suboptimality morally tolerable, namely »non-culpable misinformation or lack of information«. (Note again that we already encountered this general idea above regarding EPISTEMIC LIMES in the context of Broad's Property, cf Section [3.5.1].)

At this point, I do not want to take a position on whether Pinkert's claim is ultimately convincing. Even if it isn't, a variant of the NO-DIFFERENCE CHALLENGE certainly remains for MSAC theories (cf. Singer [1980]; Kagan [2011]; Hedden [2020]). However, even given that I explicitly excluded MSAC from the scope of this thesis, the following observation can still be drawn: there are candidates for qualified formulations of MH that may well be attractive for consequentialist theories, even for non-objective varieties. Thus, should it turn out (and it will!) that only such a softened variant would be defensible for MOAC as well, this would be tolerable – provided there were

---

seems quite reasonable that decision-theoretic, prospective subjective theories have prevailed with in the subjectivist's camp.

reasons for such a softening. To borrow from Pinkert's formulation: a moral theory that accommodates CONGRUENCE more is, other things being equal, strongly preferable to a moral theory that accommodates it less.

Even though every author has put the idea slightly differently, all of them apparently try to capture the same basic intuition, and PMH, coined by Feldman, remains the most widely accepted name for it. Accordingly, I use PMH as a name for the general idea that connects all the specific phrases. In contrast, I use COLLECTIVELY MAXIMIZING for reference to the property and MH to denote the specific criterion I introduced in the introduction and that I will use in the context of this work, i.e.:

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

My formulations are deliberately chosen so that, unlike Regan's or Portmore's, they do not include an unbound variable – just as Feldmann's and Pinkert's formulations do not. The additional generality that may be lost as a result does not matter for the purposes of the ARGUMENT. Moreover, unlike Portmore (Portmore 2018, p. 13), I propose that PMH should merely be understood as a necessary, not as well as a sufficient criterion of adequacy. Otherwise, one arguably would be committed to accepting (in that sense) trivially adequate but generally rejected moral theories, and probably would also be committed to an intolerable moral pluralism.<sup>72</sup>

---

<sup>72</sup> Assume we were to accept Portmore's version. Recall

**Criterion 3.1 (MH (Portmore))** *A moral theory T is correct if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could together bring about.*

To argue in favor of the truth of  $P_{MOCOR}$ , MH's formulation is sufficient.

To see this, we rewrite it<sup>73</sup> to:

**Criterion 1.3 (MORAL HARMONY (MH, tentative, contraposition))**

*If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome (i.e., according to this theory), (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).*

This, in turn, arguably yields:

---

There is a reading of this criterion that boils down to

- (9) A moral theory is correct if and only if [if all agents act right according to that theory, then, necessarily, they produce the morally best outcome that they could together bring about].

According to such a criterion, every theory that makes true the conditional on the right-hand side of (9), i.e.,

- (10) If all agents act right according to that theory, then, necessarily, they produce the morally best outcome that they could together bring about,

would be correct.

Now consider the following, rather prohibitive moral theory:

**Definition 3.3 (VACUOUS CRITERION OF RIGHTNESS (VACCOR))** *All options are (necessarily) wrong.*

According to the standard Lewisian account, counterfactuals with impossible antecedences are vacuously true, in analogy to material conditionals with actually false antecedences. (In the second chapter of his work on counterfactuals, David Lewis accepts this view as a result of his semantic account but also briefly discusses alternative truth conditions (D. Lewis 1973). For a more detailed assessment of different approaches to this problem of »counterpossibles« and their pros and cons, see D. H. Cohen 1987 and more recently Ferreira and Berto 2018.) Lewis' standard account certainly suffices to motivate the issue in this context.) Under this account, (10) is vacuously true and, thus, any theory embracing VACCOR is guaranteed to fulfill MH and, thus, to be correct according to (9) simply because it is then impossible for agents to perform the right actions. This seems plain wrong.

In addition, several theories might fulfill (10), most likely committing a champion of (9) to some implausible pluralism with respect to rightness.

<sup>73</sup>One merely takes the contraposition of the right-hand side of Criterion 1.2 and (arguably plausibly) reads the »guaranteed« as indicating a strict conditional. Furthermore, I translate the logically guaranteed »not act rightly« directly to »act wrongly« (but keep in mind the two possible MOAC wrongness predicates discussed at the end of Chapter 2 – I come back to this later). As indicated in Footnote 8, this translation is actually less innocent than one might initially think in light of the CONSEQUENTIALIST STANDARD VIEW. However, it certainly fits Pinkert's ON-THE-HOOK and can be justified (which I will do later in part 2 when the semantic difference actually matters).

(PMH) If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

The preceding textual work, mind you, has not put forward any further substantive arguments for PMH beyond the justification already formulated in the introduction that CONGRUENCE allows for both an individual (most notably, MOCOR) and a collective reading (PMH). However, it *should* have shown that the intuitive relevance and persuasiveness of PMH were and still are shared in vast parts of philosophical discourse. This should be enough for us to accept the premise as set for now even though, regarding its deeper semantic structure, the seemingly straightforward statement of MH is more complex than it initially appears. These nuances will be explored later in this thesis, where working these out in a precise and concise way will prepare the ground for the second, reconstructive part of this project. For now, however, we join the obviously widespread view elaborated here and next look at how, with the help of MOCOR, an alleged inner-theoretical inconsistency is derived via MOCOR (or similar explications of MOAC's criterion of rightness) from the existence of TROUBLEMAKERS against the background of PMH.

### 3.5.2.3 Step 3: Regan's ›Proof‹ and P<sub>MOCOR</sub>

The second of the »two particular intuitions [that] can not be reconciled« (see above) and thus the second puzzle piece of Regan's main argument is (cf. Regan 1980, pp. 3-4, p. 6):

**Property 3.3 (PROPAU (Regan))** *For any agent, in any choice situation, if the agent satisfies T in that situation, he produces by his act the best consequences he can possibly produce in that situation.*

From PROPAU in combination with PROPCOP Regan infers a third property (Regan [1980] p. 6):

**Property 3.4 (ADAPTABILITY (Regan))** *A theory T is adaptable if and only if the agents who satisfy T, whoever and however numerous they may be, are guaranteed to produce the best consequences possible as a group, given the behavior of everyone else.*

Regan claims that ADAPTABILITY is a »generalization of PROPAU and PROPCOP« that »is stronger than the conjunction of PROPAU and PROPCOP« (*ibid.*, pp. 6-7), which Regan proves by showing that ADAPTABILITY implies both PROPCOP and PROPAU (cf. *ibid.*, pp. 107-108). For Regan, ADAPTABILITY is the ultimate desideratum of exclusively act-oriented theories in general and, hence, of objective consequentialist theory more specifically. Finally, Regan's ultimate goal is to prove an impossibility, namely that *no exclusively act-oriented theory can be adaptable*.

Regan does so by allegedly showing that there are cases – like WHIFF AND POOF and TWO FACTORIES, i.e., MUTUAL SANCTIFICATION CASES – in which actions that are right according to such a theory lead to a violation of PROPCOP (and, thus, of MH), i.e., to suboptimal outcomes. Establishing such assessments is, thus, the third and final building block of Regan's argument and, thus, of the CHALLENGE as the CHALLENGE<sub>int</sub>.

Some authors apparently think that such assessments are so trivial that they do not give an explicit argument for them. For instance, recall that

Pinkert claimed that »Act Consequentialism judges that both Ann and Ben act rightly by polluting« (Pinkert [2015], p. 975) which is simply based on the apparent observation that »no one could have brought about better results by acting differently« (*ibid.*, p. 972). But the (often implicit) argument behind these inferences is actually quite interesting and informative when put under the microscope. Regan is explicit in this regard.

He starts his reasoning from his »precise necessary condition for exclusive act-orientation« (Regan [1980], p.114). Recall his »the partial definition«:

Any exclusively act-oriented theory must, in this example, on any assumption about Poof's (Whiff's) behavior, identify some non-empty subset of the set of acts comprising »pushing« and »not-pushing« such that Whiff (Poof) satisfies the theory if and only if he does some act from that subset. This necessary condition for a theory's being exclusively act-oriented I shall refer to, for expository convenience, as the 'partial definition' of exclusive act-orientation.

Closer inspection reveals that there are two characteristics of exclusively act-oriented theories. First, for an exclusively act-oriented theory, what makes an action right or wrong is *solely* based on features of the action itself (such as its consequences, compliance with specific rules, etc.), irrespective of, for instance, why the agent performed the action or what kind of person the agent is. Thus, examples of theories that are *not* exclusively act-oriented are all kinds of motive-based theories, such as virtue ethics or some versions of deontological ethics, which argue that the morality of an action is, at least in part, a function of the agent's motives, intentions, or character traits (cf Section [2.2]).

Concerning this first characteristic, it is rather obvious that MOAC theories are exclusively act-oriented (or, at least, fulfill Regan's necessary condi-

tion) by definition. Recall the formal definition of objective consequentialist theories:

**Definition 2.3 (Objective Consequentialist Theory (formal))** *T is an objective consequentialist theory if and only if it embraces an axiological sub-theory  $T_{Ax}$  with a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and an objective consequentialist criterion of rightness  $T_{CoR}$  such that, for all decision situations  $D \in \mathcal{I}$  and for all  $\phi \in \Phi_D : D, C \models_T R\phi$  if and only if  $T_{CoR}(\phi)$ .*

A criterion of rightness  $T_{CoR}$  is objective consequentialist if and only if, for all  $D \in \mathcal{I}$  with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  (with  $D$ 's actual context  $C$ )  $T_{CoR}$  corresponds to a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :

$$D, C \models_T R\phi \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))).$$

Further, recall that MOAC theories qualified as *objective* by definition as their characteristic function can be explicated as

$$\chi_{MOCOR, D, C}^{\text{Val}}(\text{Val}(\text{Out}_{D,C}(\phi))) = \begin{cases} \top & \text{if } \phi \in \arg \max_{\phi' \in \Phi_D} \text{Val}(\text{Out}_{D,C}(\phi')), \\ \perp & \text{otherwise} \end{cases}$$

MOCOR's characteristic function obviously operates solely on the features of the options (more precisely on their consequences) and, thus, fulfills this first condition of Regan's partial definition.

Second, and easily overlooked, Regan's definition presupposes that the →identified subsets← – representing permissible or right actions as defined by the theory under consideration – are *non-empty*. This just means that for every decision situation, there exists at least one morally right option.<sup>74</sup> We

---

<sup>74</sup>Note that this assumption prevents, for instance, VACCOR from qualifying as being exclusively act-oriented (cf. Footnote 72).

should articulate this second characteristic as a specific property of moral theories:<sup>75</sup>

**Property 3.5 (RESOLVABILITY)** *A moral theory is resolvable if and only if, for all decision situations and all relevant contexts, at least one option is right.*

RESOLVABILITY entails two other pertinent properties:<sup>76</sup>

**Property 3.6 (WEAK DEONTIC COMPLETENESS)** *A moral theory is weakly deontically complete if and only if, for all decision situations and all relevant contexts, at least one action has a deontic status.*

and

**Property 3.7 (NO MORAL DILEMMAS)** *A moral theory is free of moral dilemmas if and only if, for all decision situations and all relevant contexts, not all actions are wrong.*

Given that RESOLVABILITY inherently implies both WEAK DEONTIC COMPLETENESS<sup>77</sup>

All these properties appear, at first glance, to be typical of consequentialist theories (and they later play a central role as part of a set of criteria to assess

<sup>75</sup>I assume here that Regan implicitly presupposes that with respect to arbitrary decision situations, that is, he assumes that  $I_T = I$ .

<sup>76</sup>Similar properties have been considered for decision-making procedures in general, most importantly in the context of potential voids of responsibility, cf. Braham and Hees 2011.

<sup>77</sup>The term »weak« in WEAK DEONTIC COMPLETENESS is used to differentiate it most importantly from:

**Property 3.8 (DEONTIC COMPLETENESS)** *A moral theory is deontically complete if and only if, for all decision situations and all relevant contexts, all actions have a deontic status.*

This property is stronger than WEAK DEONTIC COMPLETENESS but isn't entailed by RESOLVABILITY.

proposed solutions to the CHALLENGE, but with that I anticipate). If, for example, MOCOR tells us that exactly those actions are right and have at least as good consequences as all their alternatives, it may seem clear that *at least* one action must have this property<sup>78</sup> and NO MORAL DILEMMAS, it's apt to term it as being *stronger* than the two.<sup>79</sup>

The crucial question now is, what options does this non-empty subset comprise in MUTUAL SANCTIFICATION CASES like WHIFF AND POOF and TWO FACTORIES according to MOAC (or, in Regan's terms, according to PROPAU). This question is far from being trivial because, on closer inspection, it seems under-specified *which* actions have the best consequences and, thus, would ›satisfy‹ MOAC. Recall the normal form of WHIFF AND POOF:

---

<sup>78</sup>Properties along these lines have been proposed as separating consequentialist from non-consequentialist thought, but ›consequentializable‹ (cf. Dreier 1993; Portmore 2009; Dreier 2011) theories (cf. Brown 2011 with respect to Property 3.7).

<sup>79</sup>This use of »stronger« is consistent with the conventional understanding in formal logic. Specifically, a proposition  $\alpha$  is deemed stronger than another proposition  $\beta$  if the truth of  $\alpha$  restricts the set of potential models more than  $\beta$ . Essentially,  $\alpha$  is true in a narrower set of circumstances than  $\beta$ . This means  $\alpha$  entails  $\beta$  without the converse being true. By extension, within the domain of moral theories and their properties, a property  $F$  is stronger than another property  $G$  when any moral theory satisfying  $F$  also satisfies  $G$ , but not necessarily vice versa.

Similar thoughts justify calling the two intertwined strands of the CHALLENGE, i.e., the CHALLENGE<sub>int</sub> and the No-DIFFERENCE CHALLENGE, the *stronger* and the *weaker* strand, respectively. Their interrelation is anchored in the hierarchy of their foundational assumptions. The CHALLENGE<sub>int</sub> is built on a set of assumptions that not only are more rigorous but also ›entail‹ the assumptions of the No-DIFFERENCE CHALLENGE – if not logically and maybe not even conceptually, at least intuitively: a MOAC theory violating MH arguably is counter-intuitive, at least from a consequentialist point of view (I hope this claim is sufficiently backed by my literature review at this point). This means that any theory satisfying the assumptions of the CHALLENGE<sub>int</sub> will, *arguably*, satisfy those of the No-DIFFERENCE CHALLENGE. Therefore, if the challenges presented by the CHALLENGE<sub>int</sub> are substantiated, they represent a more formidable and encompassing critique of consequentialism. Given this logical entailment and the differential implications for consequentialism, it's fitting to term the CHALLENGE<sub>int</sub> as the stronger strand and the No-DIFFERENCE CHALLENGE as the weaker strand. This terminology aids in demarcating the two while also highlighting the inherent intensity and breadth of challenges they pose for MOAC theories.

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

Let's ask what options of Whiff (and Poof, respectively) in WHIFF AND POOF are right according to MOAC. This is to ask (for  $X \in \{ \text{Whiff}, \text{Poof} \}$ ), which non-empty subset  $S_X$  of

$$\Phi_X := \{ \text{pushing}_X, \text{not-pushing}_X \}$$

corresponds to the characteristic function  $\chi_{\text{MOCOR},D,C}^{\text{Val}}$ . The right answer, it seems, is that *it depends* on what the respective other agent does. And this is what Regan's central argument ›exploits‹.

Here is Regan, unfolding his ›impossibility result‹ (Regan [1980], p. 115):

Suppose there is an adaptable theory  $T$  which satisfies the partial definition. Suppose further that Poof does not push. Since  $T$  satisfies the partial definition, there is some non-empty subset of the set of acts »pushing« and »not-pushing« such that Whiff satisfies  $T$  (while Poof does not push) if and only if he does an act from that subset. Call the subset  $S$ . We can deduce what  $S$  must be from the assumptions we have made about  $T$ . We know that Whiff satisfies  $T$  if and only if he does an act from  $S$ . So, if Whiff does an act from  $S$ , he satisfies  $T$ . Since  $T$  adaptable,  $T$  has PROPAU. That means that any agent who satisfies  $T$  produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, which include Poof's not-pushing, he must not-push. Therefore, if Whiff satisfies  $T$ , he not-pushes. Remembering what we have already established, that if Whiff does an act from  $S$ , he satisfies  $T$ , we can conclude that if Whiff does an act from  $S$ , he not-pushes. But remember

also that  $S$  is non-empty. The only non-empty set such that if Whiff does an act from that set he not-pushes is of course the set consisting of the act »not-pushing«. Therefore  $S$  consists of the act »not-pushing«. In sum, if Poof does not push, then Whiff satisfies  $T$  if and only if he (Whiff) not-pushes also.

We can reconstruct Regan's argument in terms of the formalism introduced in Chapter 2: first, we assume that there is a theory  $T$  that is adaptable and that satisfies Regan's partial definition of exclusive act-orientation. From this, we infer for Whiff, relative to  $D$  representing Whiff's individual decision situation in<sup>80</sup> WHIFF AND POOF with its actual context  $C$ :

1. Since  $T$  fulfills Regan's partial definition, there is a  $S \subseteq \Phi_D$  such that
  - a)  $|S| > 0$ , i.e.,  $S \neq \emptyset$  and
  - b) for all  $\phi \in \Phi_D : T, D, C \vDash \phi$  if and only if  $\phi \in S$ .
2. Since  $T$  is adaptable,  $T$  entails PROPAU – which we assume to be extensional equivalent to MOCOR – and, thus, for all  $\phi \in \Phi_D$ , if  $T, D, C \vDash \phi$ , then  $\phi \in \arg \max_{\phi' \in \Phi_D} \text{Val}(\text{Out}_{D,C}(\phi'))$ .<sup>81</sup>

Thus, unsurprisingly, in light of the formerly established formalism, it holds that

$$S = \arg \max_{\phi' \in \Phi_D} \text{Val}(\text{Out}_{D,C}(\phi'))$$

Regan claims that this is sufficient to derive what is right for Whiff (according to such a theory  $T$ ) given that Poof not-pushes. To assume that Poof not-pushes is to assume that  $C$  includes that Poof not-pushes. In this case, we know that

---

<sup>80</sup>The implicit assumption that there is such an individual decision fits well with our corresponding presumption, cf. Section 3.3 on page 64.

<sup>81</sup>This is exactly one direction of the condition we have defined for the predicate of MOCOR.

Whiff produces the best possible consequences only if he not-pushes. Thus, we know that

$$S = \arg \max_{\phi \in \Phi_D} \text{Val}(\text{Out}_{D,C}(\phi)) = \{ \text{not-pushing}_W \}$$

This, of course, is simply to say that Whiff not-pushing would have better consequences than Whiff pushing. Thus we know that, according to (2.),

(11) If  $T, D, C \vDash \phi_D$ , then  $\phi_D = \text{not-pushing}_D$ ,

that is, as Regan correctly states, to say that, relative to  $C$ :

(12) If Whiff satisfies  $T$ , then he not-pushes.

Equivalently, it holds, again relative to  $C$ , that:

(13) If Whiff does an act from  $S$ , he not-pushes.

Thus, either there is no act in  $S$  or, if there is, then it is not-pushing. But (1.a.) warrants that there is at least one act in  $S$ . Thus it is right for Whiff to not-push.

Trivially, we can prove the same with exchanged roles (Regan [1980], p. 115):

What we have just proved about Whiff we could also prove about Poof. Given our assumptions about  $T$ , if Whiff does not push, then Poof satisfies  $T$  if and only if he (Poof) does not push.

This leads Regan to the conclusion that if both not-push, they both do right according to  $T$  ([ibid.], p. 116):

But now, suppose that both Whiff and Poof not-push. We have demonstrated that each of them satisfies  $T$ , when the other not-pushes, if and only if he not-pushes. Therefore, when both not-push, both satisfy  $T$ .

The next step of Regan's argument cannot come as a surprise to anyone (Regan [1980], p. 116):

But then universal satisfaction of  $T$  does not guarantee the production of the best possible consequences.  $T$  does not have PROP COP. Adaptability entails PROP COP, so if  $T$  does not have PROP COP,  $T$  is not adaptable. That is a contradiction. We conclude that a theory which satisfies [sic!] the partial definition of exclusive act-orientation cannot be adaptable. QED.

We can now take a step back and try to get a better grip on what actually happens in this argument. For this, it is crucial to emphasize how strongly Regan's reasoning relies on the idea that the moral status of what is right for Whiff and Poof must be carefully assessed relative to the *exact* context. Take, for instance,

(14) It is right for Whiff not to push.

Is (14) true *relative* to the context spanned by WHIFF AND POOF alone or is it false? It seems that this is a false dichotomy. Instead, it seems plausible to understand it as underspecified or as undefined, i.e., as neither true nor false.<sup>82</sup> However, relative to WHIFF AND POOF, *plus* due to the fact that

#### **Fact 3.4** *Poof does not push*

it seems that (14) is true. At least, this is what Regan's argument above apparently warrants.

---

<sup>82</sup>This is not particularly spectacular. Jan Łukasiewicz, Clarence Irving Lewis, and others have drawn their motivation for the development of three-valued logic from insufficiently specified propositions (cf. C. I. Lewis [1932], McCall [1973]).

Does this mean that we *cannot say anything* about what's right in WHIFF AND POOF? Not quite, because *conditional* assessments can be derived from such considerations. For instance, it seems plausible to accept that

- (15) If Poof does not push, then it is right for Whiff not to push

is true relative to WHIFF AND POOF.

Note that it remains unsettled for now whether we should best understand such conditional assessments as indicating that the options of Whiff and Poof have, relative to WHIFF AND POOF (that is, without any further assumptions about what the other agent does), *genuine conditional* deontic status or rather *no* deontic status at all. While I want to emphasize that both understandings are at odds with what I called the CONSEQUENTIALIST STANDARD VIEW (cf. Section 2.2, page 30), we do not need to take a stand on that question for now (but I will revisit the question later).

It is helpful to have a compact notion for the concepts used here. Let  $D$  be some collective decision situation,  $C$  be the actual context,  $F$  some state of affairs (usually with respect to some of the options of the agents within  $D$ ), and let  $f$  denote the proposition that states that  $F$  obtains. For the purposes of this project,  $F$  is typically about what other agents in  $D$  (i.e., agents different from  $A$ ) will do, do, or have done. Further, to avoid unnecessary verbosity, let it be agreed that » $\llbracket C \oplus F \rrbracket$ « is syntactic sugar for » $C$  together with the assumption that  $F$  obtains«. In other words, the actual context (or ›the circumstances‹ in Regan's terms) of decision situation  $D$  is *extended* by  $F$ . We can then pin down the technique of *conditionalization* in the form of the following principle:

**Principle 3.10 (CONDITIONALIZATION)** *Let  $C$  be a context, and let  $F$  be some state of affairs. If it is true, relative to  $\llbracket C \oplus F \rrbracket$ , that  $p$ , then it is true, relative to  $C$ , that [if  $f$ , then  $p$ ].*

CONDITIONALIZATION suggests an inference scheme. For some (collective) decision situation  $D$  with context  $C$  and some state of affairs  $F$  and some propositional variable  $\varphi$ , it holds that

$$\text{if } T, D, \llbracket C \oplus F \rrbracket \vDash \varphi, \text{ then } T, D, C \vDash f \rightarrow \varphi.$$

Precisely speaking, the first half of Regan's argument, thus, established nothing else than the truth of

(16) Whiff does right if and only if Whiff does also not push.

relative to  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact 3.4} \rrbracket$  (and  $T$ ). Therefore, according to Principle 3.10, he also established that

(17) If Poof does not push, then Whiff does right if and only if Whiff does also not push.

relative to WHIFF AND POOF (and  $T$ ).

In the same way, we certainly can introduce

**Fact 3.5** *Whiff does not push.*

and derive both that it is true that

(18) Poof does right if and only if Poof does also not push.

relative to  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact 3.5} \rrbracket$  (and  $T$ ). Therefore, according to Principle 3.10, he also established that

(19) If Whiff does not push, then Poof does right if and only if Poof does also not push.

relative to WHIFF AND POOF (and  $T$ ).

As soon as (17) and (19) have been established as true (relative to WHIFF AND POOF and  $T$ ), Regan's argument proceeds with the consideration of

**Fact 3.6** Poof does not push, and Whiff does not push.

Regan claims that from the truth of (17) and (19), we can infer

(20) Poof does right, and Whiff does right.

to be true relative to  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact 3.6} \rrbracket$  (and  $T$ ).

It is, then, tempting to read Regan's argument simply like this:

**Argument:** Licensing Not-Pushing with Regan (naïve)

$P_{P \rightarrow R(W)}^{\text{Regan}}$ : If Poof does not push, then Whiff does right if and only if

Whiff does not push.

$P_{W \rightarrow R(P)}^{\text{Regan}}$ : If Whiff does not push, then Poof does right if and only if

Poof does not push.

$P_{W \wedge P}^{\text{Regan}}$ : Whiff does not push and Poof does not push.

---

$C_{R(W) \wedge R(P)}^{\text{Regan}}$ : Poof does right and Whiff does right.

This reconstruction, however, might be guilty of glossing over important semantic aspects of Regan's argument. For instance, one could (and maybe should) ask what the *frame of reference* of this argument is supposed to be: the first two premises are established relative to WHIFF AND POOF, but the third premise is false relative to WHIFF AND POOF and only true relative to  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact 3.6} \rrbracket$ . For  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact 3.6} \rrbracket$ , however, we have not established the first two premises. This makes it hard to properly assess this naïve reconstruction semantically.

An alternative and probably better way of putting Regan's argument, then, is to reconstruct it in terms of some kind of meta-language:

**Argument:** Licensing Not-Pushing with Regan (meta)

$P_{P \rightarrow R(W)}^{\text{Regan, meta}}$ : In WHIFF AND POOF: If Poof does not push, then Whiff does right if and only if Whiff does not push.

$P_{W \rightarrow R(P)}^{\text{Regan, meta}}$ : In WHIFF AND POOF: If Whiff does not push, then Poof does right if and only if Poof does not push.

---

$C_{R(W) \wedge R(P)}^{\text{Regan, meta}}$ : In  $\llbracket \text{WHIFF AND POOF} \oplus \text{Fact } 3.6 \rrbracket$ : Poof does right and Whiff does right.

I think this way of stating the argument is proper, but it seems as if some bridging principle is missing that allows us to combine and extend contexts in the required way. However, to the best of my knowledge, Regan's reasoning has never been challenged. On the contrary, as we will see below, it is repeated in a similar form to this day. Therefore, in the spirit of the reconstructive endeavor of this chapter, I will devote myself briefly to those similar argumentations and postpone a more detailed examination of the above argument with respect to its soundness until the second part of this thesis.

In the literature on the CHALLENGE, the first reconstruction is quite common. Recall David Estlund and his TRILEMMA from the beginning of this chapter. He described a structurally similar case (Estlund [2017], p. 53):

**Case 3.9 (Dr. Slice and Dr. Patch)** *Dr. Slice is a surgeon and his colleague Dr. Patch is an expert in stitching up wounds. They are faced with a situation where a patient has a tumor. If Dr. Slice makes an incision to remove the tumor, it is necessary that Dr. Patch (or someone else) stitches up the wound afterward. If the patient is both cut and stitched, his life will be saved. However, if there is surgery without stitching or (for whatever crazy reasons) stitching*

*without surgery, the patient will have an agonizing death. If nothing happens, the patient will die but will be spared some pain.*

Here is the corresponding decision situation in normal form:

		Patch	
		go golfing	patch
Slice		go golfing	the patient dies the patient dies and suffers
		cut	the patient dies and suffers the patient survives

Now, as the story goes, on this particular occasion, both Dr. Slice and Dr. Patch have plans to go golfing. Dr. Slice will not perform the surgery and Dr. Patch will not be available to stitch up the wound. As a result, the patient's condition worsens and he eventually dies. It is easy to see that DR. SLICE AND DR. PATCH is a TROUBLEMAKER and that the actually instantiated combination of actions structurally corresponds to Whiff and Poof both not-pushing their buttons. Here is Estlund's moral assessment of that particular situation (*ibid.*, p. 53):

What Slice is required to do depends on what Patch will do. [...]

Patch ought to stitch the patient if and only if Slice will be doing the surgery (stitching is possible, but pointless and harmful if there is no wound that needs stitching). But suppose that Slice will not be doing the surgery. Patch might as well go golfing. Ought Slice to cut? Well, no, because Patch will not be there to stitch, and so the surgery will only make the patient's death more painful. Slice might as well go golfing. Neither has acted (or omitted) wrongly, despite the fact that the patient will needlessly die.

Translating a bit from prescriptive to rather descriptive deontic vocabulary, we can reconstruct Estlund's argument very much like Regan's:

**Argument:** Licensing Golfing with Estlund

$P_{S \leftrightarrow R(P)}^{\text{Estlund}}$ : It is right for Patch to stitch the patient if and only if Slice will be doing the surgery.

$P_{P \rightarrow R(S)}^{\text{Estlund}}$ : It is right for Slice to do the surgery if and only if Patch will be stitching.

$P_{P \wedge S}^{\text{Estlund}}$ : Slice will not be doing the surgery and Patch will not be stitching.

---

$C_{R(P) \wedge R(S)}^{\text{Estlund}}$ : It is right for Slice to go golfing and it is right for Stitch to go golfing.

Besides the minor difference in grammatical tense, I think it is fair to say that Estlund essentially mimics Regan's argument. The only difference seems to be the temporal perspective. Where Regan uses the present tense, Estlund considers, in simple future, what the agents *will* do.

Finally, consider Felix Pinkert, the inventor of Two FACTORIES. As mentioned earlier, we read (Pinkert 2015, pp.974):

[Two FACTORIES] becomes a challenge for Act Consequentialism only once we assume that Ann and Ben are both ›uncooperative,‹ that is, each would pollute even if the other produced cleanly. [...] In The Two Factories, it is only if both agents are uncooperative that neither could have improved matters by acting differently and that Act Consequentialism judges that both act rightly. [...]

Ann and Ben each individually producing cleanly, as the other agent would then still have polluted the river, and the livelihoods of 100 workers in the cleanly producing factory would have been destroyed. [...]

Act Consequentialism judges that both Ann and Ben act rightly by polluting, even though they together could easily have brought about much better outcomes by both producing cleanly. So if Consequentialists only have Act Consequentialism to morally apprise [sic!] a situation, then they let Ann and Ben off the hook for together producing collectively suboptimal outcomes.

We can reconstruct Pinkert's argument in different ways, but if we look for a difference relative to the former two, we find one reading that is based on a backward-looking, post-hoc description. What matters in his reasoning is the fact that they »could only have made matters worse« by acting otherwise. We might go for this reconstruction:

**Argument:** Licensing Polluting with Pinkert

$P_{A \rightarrow R(B)}^{\text{Pinkert}}$ : Given that Ann did pollute, Ben could only have made matters worse by producing cleanly.

$P_{B \rightarrow R(A)}^{\text{Pinkert}}$ : Given that Ben did pollute, Ann could only have made matters worse by producing cleanly.

$P_{A \wedge B}^{\text{Pinkert}}$ : Ann did pollute and Ben did pollute.

---

$P_{\text{BW-MOCOR}}^{\text{Pinkert}}$ : If an agent could only have made things worse by acting otherwise, then they did right.

---

$C_{R(A) \wedge R(B)}^{\text{Pinkert}}$ : Ann did right by polluting and Ben did right by polluting.

The backward-looking character of Pinkert's argument distinguishes it from the arguments of Regan and Estlund. And the worries regarding the naïve reconstruction of Regan's argument do not apply here: the first three premises are warranted by [[WHIFF AND POOF $\oplus$ Fact 3.6]], and  $P_{\text{BW-MOCOR}}^{\text{Pinkert}}$  seems to be quite plausible, too, as it just looks like an innocent backward-

looking lemma of MOCOR. We will later revisit this argument and ask whether this temporal shift really is so innocent and whether it suffices to establish the fact that all agents actually *do* right by not pushing in WHIFF AND POOF.

One additional remark on Pinkert's reasoning is imperative. Pinkert correctly emphasizes the importance of the property that he calls uncooperativeness and which, as he points out, has been called *intransigence* by Michael J. Zimmerman (Zimmerman [1996] p. 257). To see that importance, consider

**Fact 3.7 (Coop-Ben)** *Actually, if Ann were to produce cleanly, Ben would produce cleanly (but if Ann were to pollute, Ben would pollute as well).*

Given Fact 3.7 we might say that Ben is willing to ›do his share‹ if others do as well. Such a ›tit-for-tat‹-like disposition is not far-fetched, and we probably encounter corresponding attitudes quite often in everyday life. Given Ben's willingness to opt for the best outcome, however, Ann *could* bring about the best possible outcome by producing cleanly. In other words

(21) It is right for Ann to produce cleanly and wrong for her to pollute.

is true relative to  $\llbracket \text{Two FACTORIES} \oplus \text{Fact 3.7} \rrbracket$  (excluding the fact that they act in an uncooperative manner, obviously). Thus, the right and wrong thing to do sometimes does not only depend on what the other agent *does* but also what they *would* do.

Pedantically speaking, both terms – »(un)cooperativeness« and »intransigence« – are a little misleading with respect to that property: The term »(un)cooperativeness« makes us think about communication, negotiations, maybe shared goals or joint payoffs. But these are all unavailable in the cases

under consideration. Similarly, the term »intransigence« only hits the spot for specific subjective contexts where agents stubbornly *insist* on doing one thing even if they are informed that others are doing something else. But we are in an objective setting here.

This gives us reason to borrow a better label from game theory. For the rest of the book, thus, we might implicitly assume *independency of actions*: whatever agents do, they do so no matter what another agent does.<sup>83</sup> It's exactly this assumption that is built into the description of TWO FACTORIES right from the start.

We can put this a bit more precisely. Let us agree to call<sup>84</sup> a combination of actions *proper* (relative to some collective decision situation) if (and only if) there is a consequence defined for it by the description of the case at hand. We can then define

**Property 3.9 (INDEPENDENCY OF ACTION)** *Let  $\mathcal{D}$  be a collective decision situation. A combination of actions that is proper within  $\mathcal{D}$  is act-independent if and only if any unilateral deviation of the combination is also proper in  $\mathcal{D}$ .  $\mathcal{D}$  is act-independent if and only if all combinations of actions are act-independent.*

The idea here is that when all unilateral deviations from proper combinations are also proper, this rules out the relevant kind of dependency involved in Fact 3.7, i.e., that an agent would change their course of action given that some agent acts in a specific way. Thus, let us assume for the remainder of

---

<sup>83</sup>The notion corresponds to the typical assumption for *non-cooperative games*, cf. John Nash (1951, p. 286): »Our theory, in contradistinction, is based on the absence of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any of the others.«

<sup>84</sup>This rough-and-ready formulation will be explicated more precisely and formally later, but for now, it does the job.

this book that all cases under consideration are act-independent.<sup>85</sup>

Let me summarize the central takeaway message from this subsection. The first part of Regan's argument was meant to establish the last premise of what I called the ARGUMENT, viz.:

( $P_{MOCOr}$ ) If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

If Regan and his successors are correct, then the agents in TROUBLEMAKERS, when performing the troublesome combinations, create, through their actions, for each other the contexts (or circumstances) sufficient to *mutually sanctify* each other's actions.

Regan's reasoning is detailed, and the passages and arguments collected and highlighted here emphasize that this aspect of Regan's conclusions resonates in contemporary philosophical discussions. This reaches far beyond the already cited examples. Derek Parfit (both in 1984 and 1988), Frank Jackson (1987), Michael J. Zimmerman (1996), and more recently, Shelly Kagan

---

<sup>85</sup> Strictly speaking, this assumption is stronger than necessary as it does not only exclude dependencies that would break the troublesomeness of some combinations of actions but *all* kinds of such dependencies. For instance, consider replacing the explicit independence assumption within the case description of Two FACTORIES with the fact

**Fact 3.8 (Contingent Anti-Ben)** *Actually, if Ann were to produce cleanly, Ben would pollute (but if Ann were to pollute, Ben would decide independently).*

This dependency certainly would not break the troublesomeness of the combination that consists of both agents polluting even though it would not align with INDEPENDENCY OF ACTION (because, in light of Fact 3.8 the combination of both producing cleanly is not proper in the relevant sense even though Ann polluting and Ben producing cleanly is). But since these dependencies only add complexity to TROUBLEMAKERS without raising new related challenges – besides, maybe, modeling issues – we can ignore them in the context of this investigation and proceed without further specifying or restricting INDEPENDENCY OF ACTION.

(2011) and Douglas Portmore (2018) have argued in similar terms, and many more examples can be found. That all agents do the right thing in the eyes of MOAC when realizing a troublesome combination of actions is more or less taken for granted in the literature.

To my knowledge, the validity of this Regan-like reasoning was never really contested. There is a passage from Fred Feldman that puts the finger on the wound, though (Feldman 1980, p. 177; I have modified Feldman's original example to match Regan's WHIFF AND POOF case):

It is agreed that Whiff ought not to push if Poof does not push. It is also agreed that Poof does not push. However, from these two premises, we may not infer that Whiff absolutely ought not to push. We cannot detach an absolute obligation from a conditional obligation and its condition.

I think this is a *crucial* point (and will come back to this later). But Feldman thought that it is a minor one because he believed that, ultimately, we can detach the non-conditional obligation (*ibid.*):

Yet in the case at hand I believe we have another premise that enables us to detach our conclusion. That premise is that the condition is ›inevitable.‹ More precisely, it is that no matter what Whiff does, Poof will not push. Nothing Whiff can do will make Poof push it. This fact, together with the conditional obligation not to push if Poof does not, entails that Whiff absolutely ought not to push.

In other words, Feldman just highlighted the relevance of INDEPENDENCY OF ACTION again.

Such easygoing, simple arguments, thus, have convinced the philosophical community for quite some time and should suffice for now in the context of this project – even though we will return to these arguments later.

In the second part of his argument, then, Regan and his followers derived the inconsistency that threatens to haunt MOAC. That is, Regan's argument can, indeed, be straightforwardly translated into the CHALLENGE as CHALLENGE<sub>int</sub> in terms of the ARGUMENT. Recall

**Argument:** The ARGUMENT (tentative)

*P<sub>3T</sub>*: There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

*P<sub>MOCoR</sub>*: If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

*P<sub>MH</sub>*: If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

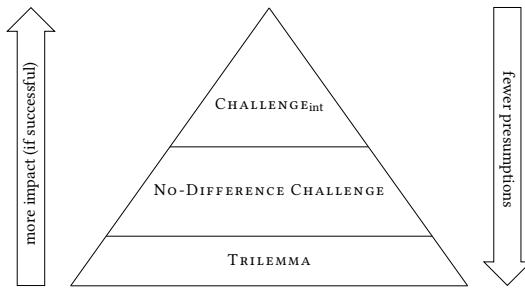
---

*C<sub>¬Adeq</sub>*: MOAC is not an adequate moral theory.

Before I continue by turning to previous attempts to solve the CHALLENGE and what would characterize acceptable solutions in terms of this project, let's have a short look back at the overall reconstruction of the CHALLENGE.

### 3.6 The PYRAMID and the Next Steps

In the opening of this chapter, I introduced the concept of the CHALLENGE as not a single variant but a hierarchy of variants, which I denoted as the PYRAMID (refer to Figure 3.12 for a recap). These formulations of the CHALLENGE differ in severity and prerequisite-richness. Lower-ranked variants within the PYRAMID are less severe, i.e., would have a less severe impact in case of success, but require weaker or fewer assumptions. Consequently, they have broader applicability, targeting not only objective but also subjective consequentialist theories or challenging even moral theories in general. Thus, lower-ranked variants can back up the higher ones in the PYRAMID.



**Figure 3.12:** The three variants that make up the PYRAMID as tackled in this thesis, ordered by their strength and dependence on the preconditions they presuppose.

At the apex of the PYRAMID, we find the CHALLENGE as CHALLENGE<sub>int</sub>, Regan's ›impossibility result‹. Should this variant be validated, MOAC must be deemed to be internally inconsistent. This is contingent on the validity of MH. Next in line is the CHALLENGE as the No-DIFFERENCE CHALLENGE, which, if successful, only convicts MOAC of extensional inadequacy. This version does not rely on the MH assumption and is relevant to all DIFFERENCE-MAKING VIEWS. At the foundation of the PYRAMID, we find the highly generalized TRILEMMA.

I hope to have elucidated the claim that the PYRAMID, compared to individual variants present in the literature, is a stronger manifestation of

the CHALLENGE. I have striven for precise representation of and sound anchoring within the literature. A comprehensive and satisfying resolution to the CHALLENGE, from the perspective of MOAC, would solve all three variants while maintaining the theory's integrity.

Of course, this does not say much about what constitutes a *good* solution. In the following chapter, I suggest such criteria and briefly offer some reasons why there is no satisfactory solution to the CHALLENGE so far.

# **Chapter 4**

## **Good Solutions, Bad Solutions, Non-Solutions**

As I delve into the fourth chapter, my focus shifts to proposed and potential solutions to the CHALLENGE. I begin by introducing what I call *solution spaces*, explicit structures that allow for taxonomizing different approaches to the CHALLENGE and tackling it systematically. This construct will serve as a sort of compass for navigating the remainder of this project as it defines the paths solutions to the CHALLENGE might follow. Further, in the first application of these solution spaces, I use them to justify the exclusion of CUMULATIVE EFFECTS CASES from the scope of my project.

Next, I turn to the question of what makes a solution actually a good one. For this, I lay the foundation for evaluating and assessing approaches to the CHALLENGE by establishing a set of specific criteria. These serve as benchmarks for the merit of different approaches and will guide my assessment of proposed solutions, providing a means to measure their effectiveness within the consequentialist framework.

Subsequently, I will examine three carefully selected, particularly influential, or especially intriguing concrete approaches. Positioned within the

solution spaces and scrutinized against the introduced criteria, I then explore whether these proposals can provide satisfactory solutions to the CHALLENGE from the perspective of MOAC. Not surprisingly, my investigation concludes that none of the solutions proposed so far is satisfactory – a new approach is needed. This concludes the first part of my work, and we move on to the second part, where I develop such an approach.

## 4.1 Mapping the Solution Space

Before delving into what it means to devise a *good* solution to the CHALLENGE within the scope of this project, it's essential to clarify what constitutes *a* solution in the first place.

To do so, let's refer back to the previously established PYRAMID (cf. Figure 4.1) and momentarily set aside any consequentialist constraints. I call a *solution* to the CHALLENGE a theory that can master all these variants. A solution should start with discharging the most dangerous variant and then continue with the next. Hence, a solution can be imagined as a path that leads from top to bottom, i.e., *out of* the CHALLENGE as CHALLENGE<sub>int</sub>, safely *through* the CHALLENGE as No-DIFFERENCE CHALLENGE, while simultaneously offering a way to *break out* of the TRILEMMA.

Thus, the first objective for a solution to the CHALLENGE is to identify a way around the most fundamental yet perilous variant – the CHALLENGE as CHALLENGE<sub>int</sub>. This variant finds its expression in what I called the ARGUMENT. Here is the general version as a reminder:

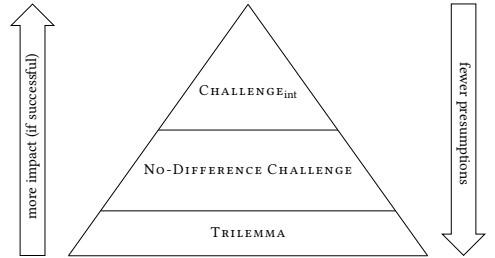


Figure 4.1: The three variants that comprise the CHALLENGE as the PYRAMID tackled in this thesis.

**Argument:** The ARGUMENT (tentative, generic)

$P_{\exists T}$ : There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

$P_{MOCOr}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{MH}$ : If  $T$  is an adequate moral theory, then (necessarily) if the agents in a collective decision situation were to act in such a way that together they would produce a morally suboptimal outcome, then at least one of them would not act morally right.

---

$C_{\neg AdeqT}$ :  $T$  is not an adequate moral theory.

One can respond to the ARGUMENT in two principal ways. First, one can try to *defend* the attacked theory. Second, typically, after failing with that first way, one can focus on *managing the defeat*. Here are more details from the specific perspective of MOAC:

1. **Defense:** If champions of MOAC want to defend against the ARGUMENT, they need to show that it is not sound. There are, as always, two lines of defense:

**Attack Validity:** Show that something is *structurally* wrong with the ARGUMENT and that it is, thus, invalid. This probably requires scrutinizing the logical-conceptual structure of the argument.

**Attack Truthfulness of a Premise:** Show that something is *substantially* wrong with the ARGUMENT by rejecting one of the premises. So there are three points of leverage in this case:

**Refute  $P_{\exists T}$ :** Usually, an attempt is made to justify  $P_{\exists T}$  by giving an example TROUBLEMAKER (or class of TROUBLEMAKERS motivated by one or more examples), which is typically given by a specific description. So, to refute  $P_{\exists T}$ , one has to show that there is something wrong with the description (or a whole class of descriptions). There are two specific manifestations of this strategy:

**Challenging the Description:** Raising doubt as to whether the supplied description satisfies the formerly introduced conditions, i.e., showing that there are reasons to believe that it is not coherent, consistent, or sufficiently complete (cf. Section 2.1, page 25).

**Challenging the TROUBLEMAKER-hood:** Raising doubt as to whether a given description actually describes a TROUBLEMAKER, i.e., showing that there actually is no troublesome combination.

**Refute  $P_{MOCoR}$ :** To refute  $P_{MOCoR}$ , one must show that the application of MOCoR identifies at least one wrong action within any troublesome combination.

**Refute  $P_{MH}$ :** To refute  $P_{MH}$ , one must show that MH is no persuasive criterion of adequacy.

2. **Manage the Defeat:** If champions of MOAC fail to defend against the ARGUMENT along the above-sketched lines of defense, they can

still try to make the best of the impending defeat. In principle, there are at least two paths to choose from:

**Biting the Bullets:** In principle, there is always the possibility that the side under attack may be prepared to live with the conclusion after all. However, with regard to the CHALLENGE<sub>int</sub>, we can exclude the option of bullet-biting for camp MOAC: a champion of a theory that accepts the *inadequacy* derived from a diagnosed inconsistency within said theory seems to violate rational standards. No one should support a theory whose theoretical inconsistency they must admit.<sup>86</sup> To ›bite‹  $C_{\neg \text{adeq}}$  as a result of the ARGUMENT would amount to just that for the followers of MOAC. Therefore, this approach cannot help with the CHALLENGE<sub>int</sub>.

**Modify & Adapt:** Alternatively, one could try to modify one's theory to allow one to defend it against the ARGUMENT without incurring *new* problems. The aim is to keep the spirit of one's theory but change details, carry out extensions, or make conditionalizations. In the context of this project, of course, particular care must be taken to ensure that the modified theory remains a MOAC theory.

This leaves us with quite a number of alternatives for responding to the CHALLENGE<sub>int</sub>. What we see in Figure 4.2 is a structured set of pathways,

---

<sup>86</sup>Note that *supporting* a theory is not necessarily to believe in its adequacy. For instance, one may well be a supporter, in a sense, of a theory that one thinks to be *useful* even though one believes it to be inadequate in the sense of being false. For example, one may consider Newtonian mechanics to be a valuable theory for all use cases relevant to one and support it to be taught to students worldwide, even if one has, of course, heard of general and special relativity. More importantly, every scientific theory is (at least very likely) false. We have indeed not reached the end of an imagined ideal scientific process, at the point of convergence of a Peirce-like theory of truth (cf. Peirce [1931] – and probably will never arrive there. But supporting a theory that one believes to be inadequate in the sense of being *inconsistent* is a different caliber altogether. *Ex falso sequitur quodlibet* and the principle of explosion are no friends of serious theories.

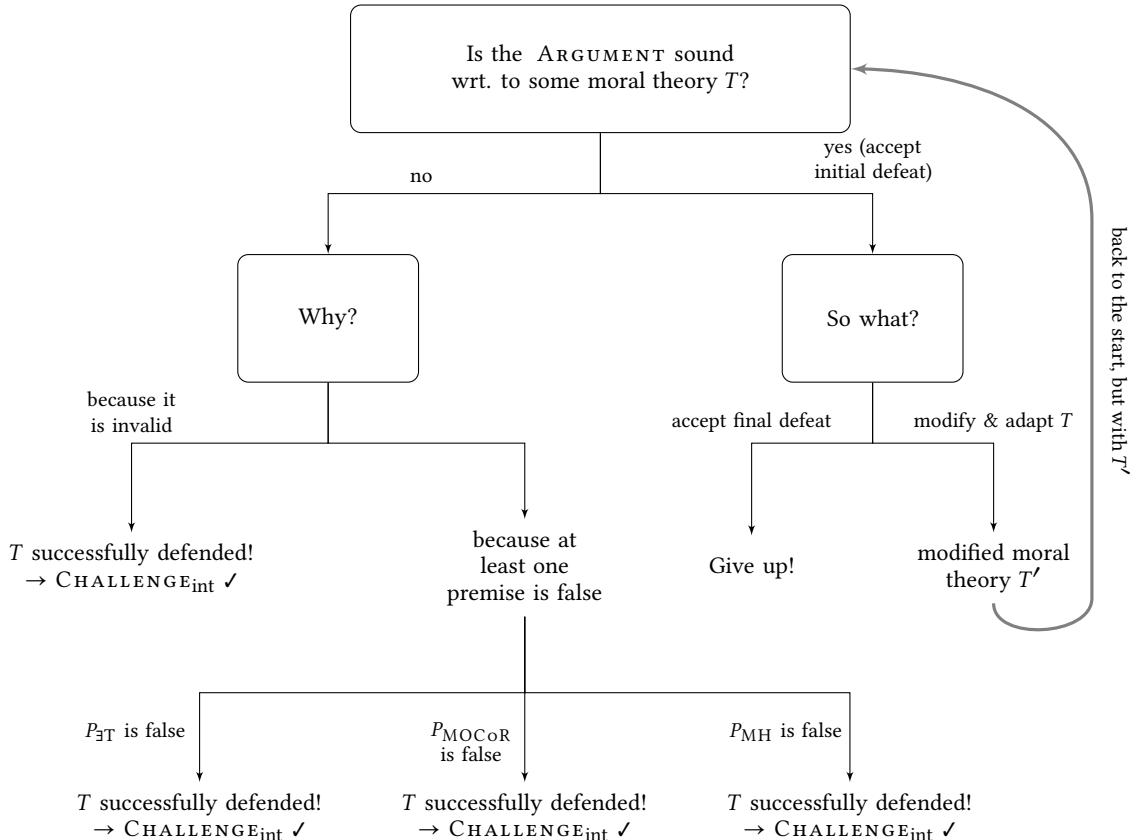


Figure 4.2: The solution space of the CHALLENGE as CHALLENGE<sub>int</sub>.

a comprehensive map which I refer to as the »solution space of the ARGUMENT« (or of the CHALLENGE<sub>int</sub>, respectively). Note that in this model, any theory that reaches one of the endpoints marked as »successfully defended« can be considered a *solution to the CHALLENGE<sub>int</sub>* (though not necessarily a satisfactory solution, especially not from the fundamental consequentialist perspective that this project adopts, we come to this later in this chapter).

It is important to remember that even once we have a solution to the CHALLENGE<sub>int</sub>, we are not done. Next, we are confronted with the two other variants of the CHALLENGE, the two other levels of the PYRAMID. The next hurdle, hence, is the NO-DIFFERENCE CHALLENGE. Recall that I also sketched an argument corresponding to that variant:

**Argument:** The No-DIFFERENCE CHALLENGE ARGUMENT (tentative)

$P_{\exists NDCs}$ : There are NO-DIFFERENCE CASES: collective decision situations in which there is at least one agent that can act in ways such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

$P_{MOCoR}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg intu}$ : If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then  $T$  is counterintuitive.

---

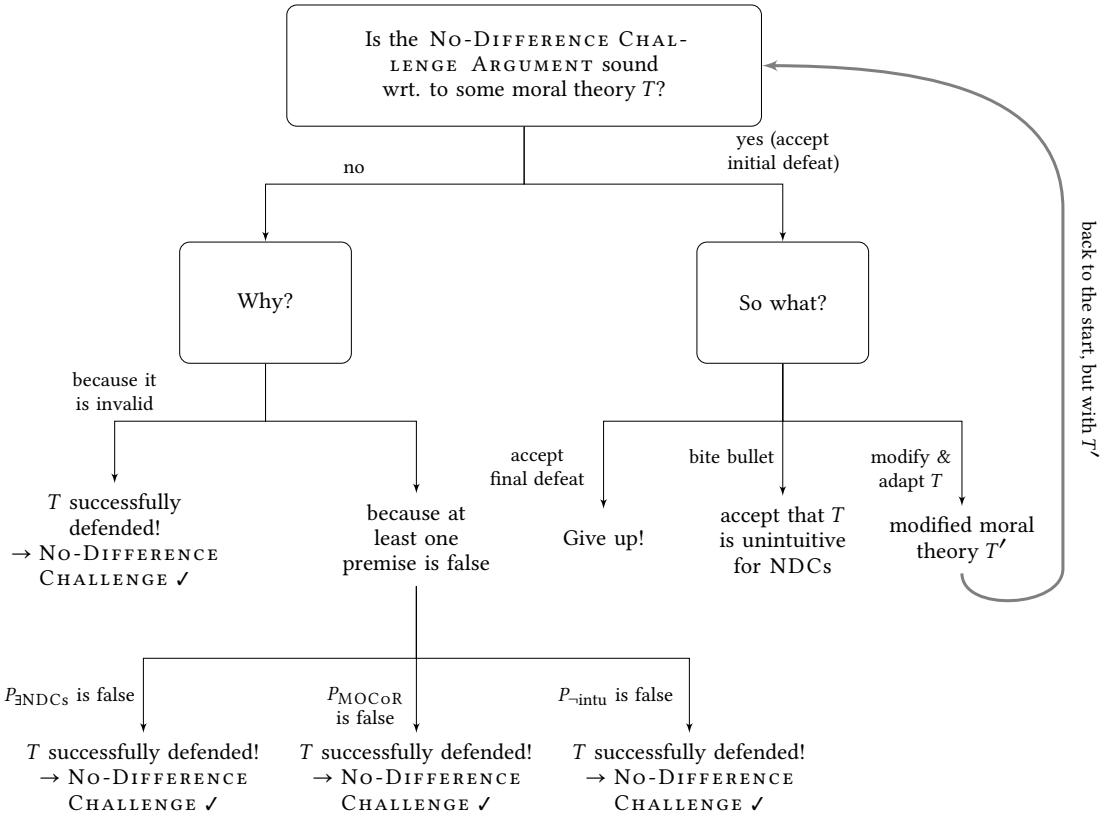
$C_{\neg intu}$ : MOAC is counterintuitive.

We have the same basic strategies available to handle the no NO-DIFFERENCE CHALLENGE. Observe that bullet-biting might very well be an option in this case. It may seem painful, but by no means intolerable, to accept a theory's unintuitiveness (like  $C_{\neg intu}$  expresses with respect to MOAC theories). Which theory does not have some counterintuitive implications, even systematic ones? Figure 4.3 shows the resulting solution space.

Finally, the TRILEMMA remains. Recall<sup>87</sup> that it arises from three apparently true propositions:

---

<sup>87</sup>The observant reader will have noticed that  $H_1$  is formulated generically at this point, i.e., it is not applied explicitly to Two FACTORIES.



**Figure 4.3:** The solution space of the CHALLENGE as No-DIFFERENCE CHALLENGE. Note that in the case of the No-DIFFERENCE CHALLENGE we may allow, in principle, for biting the bullet.

( $H_1$ ) Something wrong happens.

( $H_2$ ) If something wrong *happens*, then because someone *did* wrong.

( $H_3$ ) No one did wrong.

Since not all of these three propositions can be true, any successful defense of some theory  $T$  against the TRILEMMA needs to explain why and how at least one of them is false relative to  $T$ . If this is not possible, the options remain to give up or modify  $T$ . Figure 4.4 shows the corresponding solution space for the TRILEMMA.

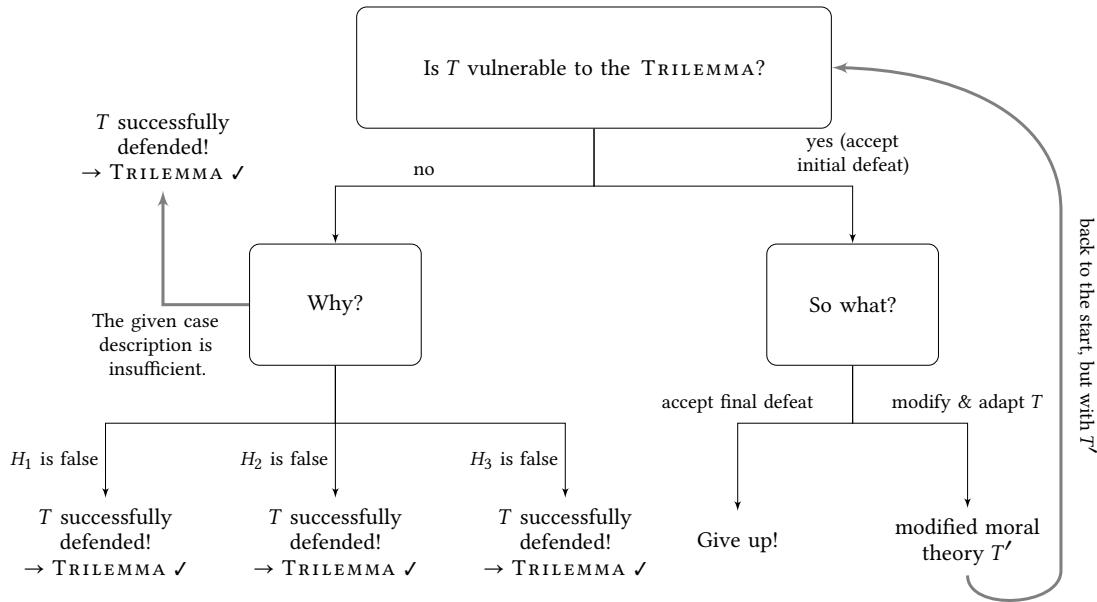
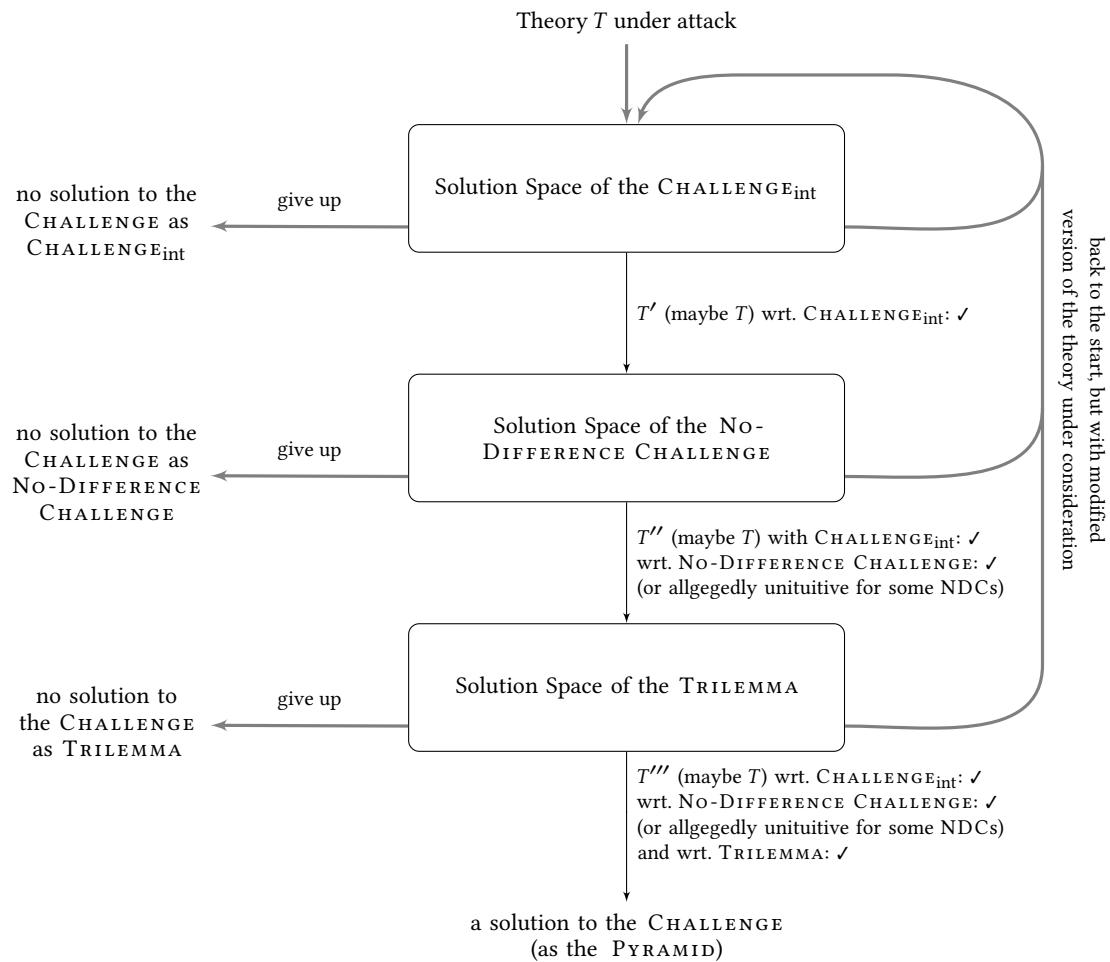


Figure 4.4: The solution space of the TRILEMMA.

Finally, we can plug all three solution spaces together to the *overall solution space* as depicted in Figure 4.5. Note that in this case, whenever we opt to modify the theory under consideration, we must make sure to start again at the very top. Otherwise, we might overlook that some modification that helps to overcome, for example, the TRILEMMA might reintroduce the ARGUMENT.

With this structural groundwork at hand, we have a precise understanding of what it means for a theory to be *a solution* for the CHALLENGE (and its subvariants). Next, we can turn to the question of what constitutes a *good* solution in terms of the consequentialist aspirations of this thesis. But before that, right at the beginning, I would like to restrict the scope of this work based on the solution space.



**Figure 4.5:** The solution space of the CHALLENGE as the PYRAMID.

## 4.2 A Limitation: On the Exclusion of CUMULATIVE EFFECTS CASES From the Scope of this Project

In the introduction, I claimed that the CHALLENGE arguably lies at the heart of some of the most pressing practical issues of our time, like the anthropogenic climate crisis. The central propositions in this context are

- (22) It would be better to reduce carbon emissions significantly.

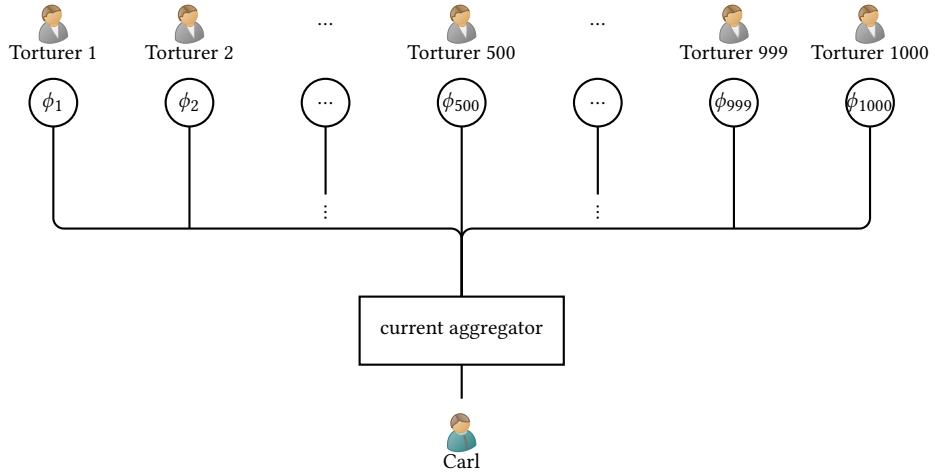
- (23) If sufficiently many people refrain from flying, become vegetarians, or switch to public transport, then this would cause a significant reduction in carbon emissions.
- (24) An individual refraining from flying, becoming vegetarian, or switching to public transport reduces carbon emissions only morally insignificantly or even not at all.

Accordingly, in the case of the climate crisis, there seems to be a myriad of CUMULATIVE EFFECTS CASES as there are undoubtedly many overlapping and not just a single potentially troublesome combination. My individual contribution to the global disaster seems as negligible as it does now, even if my neighbor were to become a bike-loving vegan. But which combinations are troublesome exactly? And under what circumstances? Are there, for example, threshold values above which my switch from car to bicycle and from beef to tofu *would* make a difference?

Furthermore, real-life cases are especially challenging as they come with several *empirical* issues. How many CO<sub>2</sub> molecules make what difference in temperature where on earth for how long? How many of them make what kind of natural disaster how much more likely? This list could probably be continued almost indefinitely. Even if, at the very least, there are statistical correlations, they make the cases less than clear. For example, somewhat simplistically, one ton of CO<sub>2</sub> emitted is associated with losing three square meters of sea ice (cf. Notz and Stroeve 2016). What does such a finding say about individual (in)effectiveness? Would it be appropriate to assign each and every agent their ›fair share‹ even if we do not find a causal link?<sup>88</sup> A

---

<sup>88</sup>This is, roughly, Glover's so-called »Share-of-the-Total-View« (cf. Glover and Scott-



**Figure 4.6:** The setup of the HARMLESS TORTURERS case: 1000 torturers can each flip their switches  $s_1$  to  $s_{1000}$ ; the number of switches flipped determines the strength of the shock Charly receives.

project like the present one, i.e., a project that aims at *theoretic* and *conceptual* challenges, is grateful if it does not have to lose itself in the subject area of empirical sciences. I thus will not, and cannot, scrutinize real-life candidates for TROUBLEMAKERS.

But even if we leave real-world cases aside and circumvent these empirical complications as far as possible, CUMULATIVE EFFECTS CASES still comes with several conceptual difficulties. Recall

**Case 3.6 (HARMLESS TORTURERS)** *Carl is wired to a torture machine with a thousand identical switches. When none of the switches are flipped, no current runs through the machine, so Carl is in no pain. If all a thousand switches are flipped, then a sizable current runs through the machine, and Carl is in tremendous pain (but no permanent damage is done to his body). But the flipping of any given switch increases the current only by a tiny amount (well below the perceptually discriminable threshold for pain) so that Carl simply cannot tell*

---

Taggart [1975]. Parfit dismissed this view on convincing grounds (cf. Parfit [1984] chapter 3, section 25, and so we do not need to consider it in this chapter in any more detail, but see Figure 4.12 for classification of the approach).

*whether one switch more or less has been flipped – regardless of how many other switches have already been flipped. Finally, imagine that a thousand different people each control a single switch and must decide whether to flip it or not. None of them cares about Carl or feels any remorse, but each of them enjoys flipping switches.*

Obviously, it would be best if sufficiently many torturers would refrain from flipping their switches – whereby »sufficiently many« is hard to impossible to define given the vague nature of cumulative effects. Thus, even though it is a carefully designed thought experiment, HARMLESS TORTURERS is by no means a simple and straightforward case. It rests on the vagueness of certain relevant predicates and the psychology of perception. Conceptually, the struggle starts already when we think about the fact that HARMLESS TORTURERS is described such that each and every flip

1. brings some joy to the torturer flipping the switch,
2. does not increase the harm at all or increases it only insignificantly (i.e., some imperceptible harm is inflicted, or whatever else characterizes insignificance here) – most importantly, whether the participation of one torturer more or less makes things not worse –,
3. and yet the sum of these (non-)contributions somehow aggregates to more harm than ›the joys of flipping switches‹.

These ›aggregative facts‹ certainly can make one scratch one's head. Shelly Kagan even claimed that such cases are *impossible* as, at some point – which might be hard or impossible to determine –, there *must* be a »perceptible difference« (Kagan 2011, p. 134, fn 13). Here is his slightly adapted *reductio ad absurdum*:

By hypothesis, when the person is in state 0, they are in no pain. If we ask them whether they are in pain, they will answer »no.« In state 1,000, they are in excruciating pain. If we ask them whether they are in pain, they will answer »yes.« Suppose then we consider state 1. Since this is adjacent to state 0, the difference between state 0 and state 1 must be imperceptible. Hence, if we ask someone in state 1 whether they are in pain, they must give the same answer as they gave when the same question is posed with regard to state 0, that is, they must answer »no.« (If their answer in state 1 differed from their answer in state 0 this would presumably indicate a difference in their perception of the two states, contrary to hypothesis.) Now consider state 2, which is, of course, adjacent to state 1. Since, by hypothesis, the two adjacent states are imperceptibly different, the answer to the question »are you in pain?« must be the same. But the answer to this question with regard to state 1 is »no,« hence the answer with regard to state 2 must be »no« as well.

First, note that Kagan considers HARMLESS TORTURERS as a SEQUENTIAL CASE, i.e., there are »adjacent« states that can be compared to each other. But this is only one of several implicit assumptions Kagan's argument rests on. For instance, it rests upon the assumption that the indiscriminability of two states implies that they are *equally* good (and not just, say, *on par*). Further, the later part of his reasoning rests upon the assumption that the overall harm is the sum of trigger harms. A slightly generalized version of Kagan's argument goes like this:<sup>89</sup>

---

<sup>89</sup>See Nefsky [2011] and especially Spiekermann [2014] and Hedden [2020] for more detailed reconstructions and critical investigations of Kagan's argument. The two latter explicate quite precisely the possibility that being morally on par does not necessarily entail being good.

**Argument:** Conceptual Impossibility of CUMULATIVE EFFECTS CASES

$P_1^{\text{Imp}}$ : If CUMULATIVE EFFECTS CASES are conceptually possible, then, for all pairs of adjacent states, the subsequent state is not morally worse than the proceeding state.

$P_2^{\text{Imp}}$ : If, for all pairs of adjacent states, the subsequent state is not morally worse than the proceeding state, then the end state cannot be morally worse than the start state.

$P_3^{\text{Imp}}$ : If CUMULATIVE EFFECTS CASES are conceptually possible, then the end state is morally worse than the start state.

---

$C^{\text{Imp}}$ : CUMULATIVE EFFECTS CASES are conceptually impossible.

Whether this (obviously conceptually valid) argument goes through or not certainly depends on whether  $P_2^{\text{Imp}}$  is true (given that  $P_1^{\text{Imp}}$  and  $P_3^{\text{Imp}}$  are true by definition or description of CUMULATIVE EFFECTS CASES).

Suppose Kagan's argument is sound (or, to be more exact, even if  $C^{\text{Imp}}$  were established on different grounds). In that case, consequentialists can rightfully ignore CUMULATIVE EFFECTS CASES with respect to their handling of the CHALLENGE – because this class of TROUBLEMAKERS simply were empty then. So, what can we say about  $P_2^{\text{Imp}}$ ?

The answer is not as easy as Kagan wants us to believe. In fact, it depends on several axiological questions and on how we interpret specific phrases that are involved – i.e., it becomes at least in part a question of modeling (Spiekermann 2014, E. N. Dzhafarov and D. D. Dzhafarov 2010a; E. N. Dzhafarov and D. D. Dzhafarov 2010b, and Hedden 2020). It seems thus possible to bend the axiological part of one's theory in such a way that Kagan's argument *fails*. However, the necessary adjustments, restrictions, and extensions, mainly re-

lated to vagueness, lead to an arguably otherwise unnecessarily complicated theory of value aggregation. Consequentialists might ask rightfully: Is it really the task of consequentialism to dig traps for itself? Is it not rather the task of the one who attacks consequentialism to at least specify such an axiology precisely and actually even to justify it as particularly plausible?

In conclusion, considering Kagan's impossibility argument, it seems reasonable for the consequentialist to call for a shift in the burden of proof. Before we proceed to further defenses, it is incumbent on those asserting the CHALLENGE to clarify precisely what the issues are. As far as I know, this has not yet been achieved successfully. Therefore, that is I exclude CUMULATIVE EFFECTS CASES from my discussion. In terms of the solution space of the CHALLENGE<sub>int</sub>, this means that I simply deny the existence of an entire class of potential TROUBLEMAKERS, namely that of CUMULATIVE EFFECTS CASES. Therefore, I reject  $P_{\exists T}$  of the ARGUMENT, I reject  $P_{\exists NDCs}$  wrt. to the NO-DIFFERENCE CHALLENGE, and for the TRILEMMA, I also dismiss the descriptions of the corresponding cases as insufficient or inconsistent. This guarantees that we have a trivial path to take through the overall solution space of the CHALLENGE in terms of the PYRAMID for CUMULATIVE EFFECTS CASES – we do not need to investigate in any further way.

### 4.3 Criteria for Good Solutions

With our understanding of the solution space established and a clear idea of what constitutes a solution, we're now ready to delve into the qualitative aspects of these solutions. We need a set of criteria to decide *what makes a solution good or bad* – or, at least, what makes one solution better or worse than another. Since, in our setting, solutions are theories, this quest boils

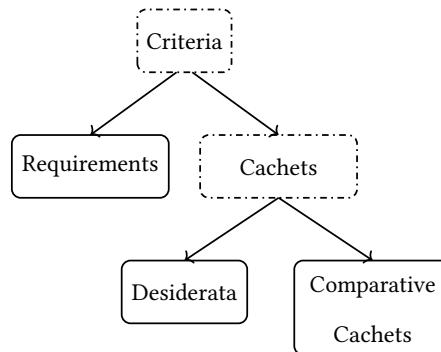
down to determining what makes better or worse theories. But let's be clear: we are not embarking here on an expedition to solve *all* of moral philosophy. Instead, our task here is more specific. We're interested in what makes one solution better than another in the context of this very project.

I introduce various criteria to evaluate and assess existing approaches to the CHALLENGE. There are three different classes of such criteria: *requirements*, *desiderata*, and *comparative cachets*. Figure 4.7 depicts the relationship between these types, and in the following, I sketch

their characteristics and offer some examples.

Along the way, I will make a selection of criteria that will be used for the assessment of solutions in the rest of this book. Mostly, I will not provide detailed arguments in their favor but rely heavily on these criteria being proposed elsewhere and intuitively convincing.

It should be noted that neither the classification nor the criteria selection is ›theoretically innocent‹. Such innocence, however, would also not fit the project's overall aim. After all, the primary goal is to solve the CHALLENGE from the perspective of MOAC theories, and thus we are aiming here at a way to rank solutions based on their effectiveness in addressing the CHALLENGE from that very perspective. Thus, our yardstick for quality is closely related to the overall success criteria of this project, i.e., developed with an eye on the formerly introduced solution spaces. As a result, the choice of the following criteria is heavily loaded with theoretical preconceptions.



**Figure 4.7:** Schematic representation of the relation between the three different criteria. All criteria fall in the leaf categories of this tree.

### 4.3.1 Requirements:

Some criteria that I will call *requirements* are *necessary conditions of adequacy*. If a theory fails to meet any of these necessary conditions, it becomes disqualified and thus is removed from consideration as a plausible solution.

Some requirements are indisputable and convincing independently of further moral background assumptions.<sup>90</sup> Consider:

**Property 4.1 (DEONTIC CONSISTENCY)** *A theory is deontically consistent if and only if there is no decision situation such that, for any relevant context, some action is both right and not right according to that theory.*

A theory permitting the same action to be right *and* not right – not ambiguously, but as a substantial claim about its moral status in the very same sense – would be deemed inconsistent (cf. Footnote 86). Similarly, consider:

**Property 4.2 (CONCEPTUAL DEONTIC CONSISTENCY)** *A theory is conceptually deontically consistent if and only if there is no decision situation such that, for any relevant context, some action is both right and wrong.*

That an action is right *and* wrong is not a strict *logical* contradiction, even though it entails one. According to the *meaning* of the involved terms, a wrong action is necessarily one that is not right. Thus, a theory allowing for a situation where the same action is right and wrong would be one that is *conceptually* inconsistent.

Another hot candidate for a more theory-independent Requirement is **METHODOLOGICAL INDIVIDUALISM**. Recall

---

<sup>90</sup>Cf. Timmons 2001, p. 11.

**Principle 2.1 (METHODOLOGICAL INDIVIDUALISM)**

*The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the act of a moral agent has this status merely in a derivative sense, that is, that deontic status is a function of the deontic status of certain options of moral agents.*

Other candidates for requirements *are* dependent on certain background assumptions. For instance, given an objective consequentialist perspective, one such requirement could be MH itself. Recall

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

However, since MH is an essential part of the CHALLENGE (as the CHALLENGE<sub>int</sub>), we should not add it to our set of requirements. Either a proposed solution to the CHALLENGE (as the CHALLENGE<sub>int</sub>) respects MH, or MH will be modified as part of it – and then this should be done for good reasons (depending on the path taken within the solution space sketched above, cf. Figure 4.2). To elevate it to the status of a necessary condition for a good solution would, therefore, unduly narrow the the space of good solutions.

However, there is another candidate for a requirement that depends on our moral philosophical background assumptions that we *should* add to the set of requirements. Since we seek objective consequentialist solutions, the theories we seek should qualify as such theories. In other words, they must fall under the definition at the end of Chapter 2:

**Definition 2.3 (Objective Consequentialist Theory (formal))** *T is an objective consequentialist theory if and only if it embraces an axiological sub-theory  $T_{Ax}$  with a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and an objective consequentialist criterion of rightness  $T_{COR}$  such that, for all decision situations  $D \in \mathcal{I}$  and for all  $\phi \in \Phi_D : D, C \vDash_T R\phi$  if and only if  $T_{COR}(\phi)$ .*

*A criterion of rightness  $T_{COR}$  is objective consequentialist if and only if, for all  $D \in \mathcal{I}$  with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  (with  $D$ 's actual context  $C$ )  $T_{COR}$  corresponds to a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :*

$$D, C \vDash_T R\phi \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))).$$

Being an objective consequentialist theory is a *necessary condition* for any acceptable solutions (relative to the purpose of this project). After all, we came here to defend MOAC – and if a proposed solution to the CHALLENGE ceases to be a MOAC theory, then this is not what we are looking for. So, we include falling under the above definition in our list of requirements.

Note that requirements divide all moral theories  $\mathcal{M}$  into two classes:<sup>91</sup> disqualified theories  $\mathcal{M}_{\text{dis}}$  and candidate theories  $\mathcal{M}_{\text{can}}$ , as

$$\mathcal{M} = \mathcal{M}_{\text{dis}} \sqcup \mathcal{M}_{\text{can}}$$

with

$$\mathcal{M}_{\text{can}} := \{ T \in \mathcal{M} \mid T \text{ fulfills all requirements} \}.$$

### 4.3.2 Cachets:

The second kind of criteria, called »cachets«, may serve as tie-breakers for candidate theories  $T \in \mathcal{M}_{\text{can}}$ . Cachets are of two kinds:

---

<sup>91</sup>For those not familiar with this notation:  $X = X_1 \sqcup X_2$  is meant to express that the elements of  $X$  are partitioned into two classes  $X_1$  and  $X_2$ , i.e., two non-empty, mutually disjoint sets. More formally:  $X = X_1 \sqcup X_2$  if and only if  $X_1 \neq \emptyset \neq X_2$  and  $X_1 \cup X_2 = X$  and  $X_1 \cap X_2 = \emptyset$ .

*Desiderata:* The cachets I call »desiderata« are properties that a theory may or may not possess. If a theory satisfies a desideratum, it gains favor but not necessarily endorsement. Desiderata help to rank candidates and allow comparisons. Consider two candidate theories,  $T_1$  and  $T_2$ . If  $T_1$  satisfies an additional desideratum over  $T_2$ , then, other things being equal,  $T_1$  is better than  $T_2$ . A plausible candidate for a desideratum is Regan's

**Property 3.5 (RESOLVABILITY)** *A moral theory is resolvable if and only if, for all decision situations and all relevant contexts, at least one option is right.*

Similarly, we might consider some of the principles RESOLVABILITY implies, like, for instance, WEAK DEONTIC COMPLETENESS (cf. page 127)

**Property 3.6 (WEAK DEONTIC COMPLETENESS)** *A moral theory is weakly deontically complete if and only if, for all decision situations and all relevant contexts, at least one action has a deontic status.*

or NO MORAL DILEMMAS,<sup>92</sup> recall (cf. page 127):

**Property 3.7 (NO MORAL DILEMMAS)** *A moral theory is free of moral dilemmas if and only if, for all decision situations and all relevant contexts, not all actions are wrong.*

Alternatively, one might opt for the stronger version of the first one, i.e., (cf.

Footnote 77 on page 127)

---

<sup>92</sup>Note that there could be, in principle, theories that have NO MORAL DILEMMAS but not RESOLVABILITY. For instance, if, *pace CONSEQUENTIALIST STANDARD VIEW* (cf. Section 2.2 page 30), some theory allows for situations with options having *no* (unconditional) moral status at all. (Note that the various variants of DEONTIC COMPLETENESS and the question of purely conditional moral status can be understood as applications of what Timmons calls »Determinancy« according to which »A moral theory should feature principles that together with relevant factual information entail determinate moral verdicts about the morality of actions, persons, and other objects of evaluation in a wide range of cases«, cf. Timmons 2001, pp. 11.)

**Property 3.8 (DEONTIC COMPLETENESS)** *A moral theory is deontically complete if and only if, for all decision situations and all relevant contexts, all actions have a deontic status.*

Whether No MORAL DILEMMAS is considered a plausible candidate for an requirement or a desideratum (if at all!) certainly depends heavily on one's moral background assumptions. Deontologists may be able to live with the fact that they are committed to the existence of moral dilemmas, e.g., Kant, with respect to the murderer at the door (but see Cholbi [2009]). However, from a consequentialist point of view, it certainly is a good candidate for a requirement (cf. Brown [2011]).

It is not to be expected that a set of plausible desiderata will induce a nice and clean total order over candidate theories. Assume there are two sets of desiderata  $\mathcal{D}_1$  and  $\mathcal{D}_2$  that at least partially distinct, i.e., with  $\mathcal{D}_1 \setminus \mathcal{D}_2 \neq \emptyset \neq \mathcal{D}_2 \setminus \mathcal{D}_1$ . What are we to do with cases where some candidate theory  $T_1$  fulfills all the desiderata in  $\mathcal{D}_1$  while some other candidate theory  $T_2$  fulfills all the desiderata in  $\mathcal{D}_2$ ? One could try to introduce some kind of hierarchy of desiderata, but I will not try to establish or defend such a hierarchy because I cannot see promising candidates for the required meta-criteria.

Instead, let us think of desiderata as inducing a *partial* order  $\prec \subseteq \mathcal{M}_{\text{can}} \times \mathcal{M}_{\text{can}}$  over the set of acceptable theories. Here is how this would work. Let  $T_1$ ,  $T_2$ , and  $T_3$  be three candidate theories, i.e., let all three fulfill all agreed requirements. Further, let  $T_3$  fulfill all desiderata in  $\mathcal{D}_1 \cup \mathcal{D}_2$ , while  $T_1$  only fulfills  $\mathcal{D}_1$  and  $T_2$  only fulfills  $\mathcal{D}_2$ . The resulting partial order, including the incommensurability of  $T_1$  and  $T_2$ , is captured in Figure 4.8.

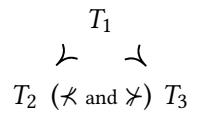


Figure 4.8: The partial ordering of the theories  $T_1$  to  $T_3$ . We can only establish that  $T_3$  is better than both  $T_2$  and  $T_3$ . However,  $T_2$  and  $T_2$  are incommensurable.

*Comparative Cachets:* Lastly, some ›criteria‹ are inherently comparative, relating theories rather than assessing them categorically. As an example, consider two traditional principles:

**Principle 4.1 (SIMPLICITY)** *All things being equal, a theory T is, ceteris paribus, superior to another theory T' if T is simpler than T'.*

**Principle 4.2 (PARSIMONY)** *All things being equal, a theory T is, ceteris paribus, superior to another theory T' if T is more parsimonious than T'.*

Although neither principle plays a particularly prominent role in this project, a few comments are in order. First, it remains somewhat vague what exactly is meant by simplicity and parsimony. As an approximation, the following two suggestions may suffice for us. A theory is simpler than another if it requires fewer or, at any rate, less complicated propositions to explain an observation or phenomenon. (In the case of moral theories, the explanandum would be, e.g., the moral status of certain options or actions.) In contrast, one theory is more parsimonious than another if it requires fewer kinds of entities to provide such explanations. Sometimes parsimony is seen as an explication of simplicity, or it is assumed that being more parsimonious implies being simpler.<sup>93</sup> Certainly, introducing additional types of entities tends to make things more complicated in the first place. But in principle, there is nothing that excludes that a theory may well provide simpler (less complicated) explanations through employing additional kinds of entities.

---

<sup>93</sup>For example, I am occasionally referred by computer scientists to a passage from a textbook on machine learning in which we read (Gori, Betti, and Melacci 2023 p. 101, my italicization): »The parsimony principle (*lex parsimoniae* in Latin) is typically connected with classic Occam razor in philosophy, which states that entities should not be multiplied beyond necessity. Hence, whenever we have different explanations of the observed data, the *simplest* one is preferable.«

Both criteria are inherently comparative in that neither simplicity nor parsimony yields useful unary predicates by themselves. Even if we had an (intuitive) understanding of what makes a theory *simple* (or *parsimonious*), it seems hard to see to what extent this says anything about the quality of the theory as such. However, if one theory is *simpler* (or *more parsimonious*) than another, *ceteris paribus*, it can be deemed better.

Now and then, it will be useful to think of extensional adequacy as a comparative cachet as well. Because, of course, we do not know for *all* decision situations which actions are right – why would we need moral theories then? – we restrict the corresponding evaluation to a *core* of »clear, trivial, obvious« cases. (There is no need to specify this set here. We can fill it later when concrete testbeds for certain theories are needed.)

#### **Principle 4.3 ((CORE) EXTENSIONAL ADEQUACY)**

Let  $I_{Core} = \left\{ \langle D, C, \Phi_D^{right} \rangle \mid D \in I_T, \Phi_D^{right} \subseteq \Phi_D \right\}$  be the core test set, where *C* is the actual context of *D* and  $\Phi_D^{right}$  is the set of actions assumed with certainty to be right in *D*. All things being equal, a theory *T* is, *ceteris paribus*, superior to another theory *T'* if *T* is extensively more adequate than *T'* wrt.  $I_{core}$ .

This formulation of (CORE) EXTENSIONAL ADEQUACY leaves unspecified what *exactly* it means for some theory to be »extensively more adequate« than another wrt. the core cases. I propose that we should simply count the correctly identified right actions according to *T*<sub>1</sub>, subtract the incorrectly assessed actions, and then do the same wrt. *T*<sub>2</sub> and compare the results.<sup>94</sup>

This informal understanding will serve my purposes well enough.<sup>95</sup>

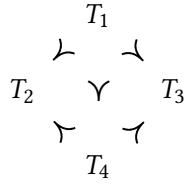
---

<sup>94</sup>For simplicity's sake, assume that all true (false) positives and true (false) negatives and all test cases  $I_{Core}$  are equally important and thus count equally. However, see the observation in Footnote 57.

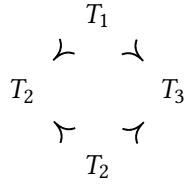
<sup>95</sup>For critical authors who demand a more precise notion, we can define:

Even if we assume that each comparative criterion might impose a total order over candidate theories, we still cannot reasonably hope that to be true with respect to the *entirety* of comparative criteria considered at once. Some theory  $T$  can be simpler than another theory  $T'$ , and yet  $T'$  might be more parsimonious than  $T$ . As with desiderata, we thus may have to live with some sort of incommensurability.

For illustration, we can reuse our example from above and add another candidate theory,  $T_4$ . Let  $T_1$  be simpler than  $T_4$  (which is equally simple as  $T_2$  and  $T_3$ ) while otherwise,  $T_1$  and  $T_4$  are on par. Assume that there are no further cachets in the game. In this case, we can rank  $T_1$  and  $T_4$  both higher than  $T_2$  and  $T_3$  and  $T_1$  higher than  $T_4$  (cf. Figure 4.9). But now consider another comparative cachet, namely parsimony. Assume that everything is as before, but  $T_4$  is more parsimonious than



**Figure 4.9:** The partial ordering of the theories  $T_1$  to  $T_4$  and with **SIMPLICITY** only. With respect to desiderata, everything is as in the example before, but with respect to **SIMPLICITY**,  $T_1$  is simpler than  $T_4$  (which is equally simple as  $T_2$  and  $T_3$ ), while otherwise  $T_1$  and  $T_4$  are on par. This allows us to rank  $T_1$  above  $T_4$ .



**Figure 4.10:** The partial ordering of the theories  $T_1$  to  $T_4$  with **SIMPLICITY** and **PARSIMONY** as criteria. Everything is as in the example before, but  $T_4$  is more parsimonious than  $T_1$  (which is equally parsimonious as  $T_2$  and  $T_3$ ). Now,  $T_1$  cannot be ranked above  $T_4$  (and neither can  $T_4$  be ranked above  $T_1$ ). They are incommensurable.

$$\text{EARel}(T_1, T_2) = \sum_{\langle D, C, \Phi_D^{\text{right}} \rangle \in I_{\text{Core}}} \left( \frac{\text{true positives of } T_1}{|\mathcal{T}_1(D, C) \cap \Phi_D^{\text{right}}|} - \frac{\text{false positives of } T_1}{|\mathcal{T}_1(D, C) \setminus \Phi_D^{\text{right}}|} \right) - \left( \frac{\text{true positives of } T_2}{|\mathcal{T}_2(D, C) \cap \Phi_D^{\text{right}}|} - \frac{\text{false positives of } T_2}{|\mathcal{T}_2(D, C) \setminus \Phi_D^{\text{right}}|} \right)$$

In other words, then, some theory  $T_1$  is better than some other theory  $T_2$  in terms of (CORE) EXTENSIONAL ADEQUACY if and only if  $\text{EARel}(T_1, T_2) > 0$ ;  $T_2$  is better than  $T_1$  in terms of (CORE) EXTENSIONAL ADEQUACY if and only if  $\text{EARel}(T_1, T_2) < 0$ ; and the larger the absolute value, the stronger this advantage of one theory of another.

$T_1$  (which is equally parsimonious as  $T_2$  and  $T_3$ ). Now  $T_1$  has an advantage over  $T_4$  and the other way around, and whether these advantages ›cancel out‹ might be undefined. This, then, means that we cannot say that  $T_1$  is better than  $T_4$  nor vice versa – nor that they are equally good (cf. Figure 4.10).

### 4.3.3 Some Notes on Criteria

All criteria – at least all of the criteria that I am going to introduce and apply in this book – fall into the above-introduced three categories. As described earlier, requirements function as a ›filter‹ for theories, while the other two kinds of criteria help us to establish a partial order of acceptable theories. Thus, with regard to the question of whether one can find *the best theory*, a rather ›sober position‹ seems advisable.

Further, it is useful to think of requirements as being ›procedurally upstream‹ to cachets in the sense that, when considering a theory  $T$ , once we have found a requirement that  $T$  violates, we do not need to consider desiderata or comparative criteria. The latter two kinds, however, I will assume to be ›procedurally on par‹, i.e., desiderata fulfillment does not, in principle, count more than having the upper hand over some theory with respect to some comparative criterion.

As already indicated in connection with the previously mentioned examples, there is a second, orthogonal dimension along which we can assess the criteria: their scope. Some criteria are so broad that they arguably hold for theories in general. Other criteria are criteria only for normative or even only for moral theories. Finally, there are criteria that are plausible for consequentialist theories – and even some that are, if at all, plausible criteria for *specific* subvariants of it, like MOAC. Furthermore, some property might be

a desideratum for theories in general, but a requirement for a specific theory or family of theories. We have encountered such a distinction already in Section 3.5.2.2, where I argued that MOAC seems plausibly committed to MH, while MSAC is not. Figure 4.11 shows an overview of the different kinds of theories and how they relate to each other using a Venn diagram.

That being said, I have not yet taken a final position on the question of what criteria to adopt – although the selection of the previous examples was not entirely arbitrary, of course. Accordingly, as Table 4.4 shows, I will start with those criteria that I consider to be quite plausible candidates (as I have also justified in passing above and in the previous chapter), especially from MOAC's point of view, of course. Nevertheless, in the following section, I will briefly explain why I think it would be *inappropriate* to elevate DEONTIC COMPLETENESS to a requirement.

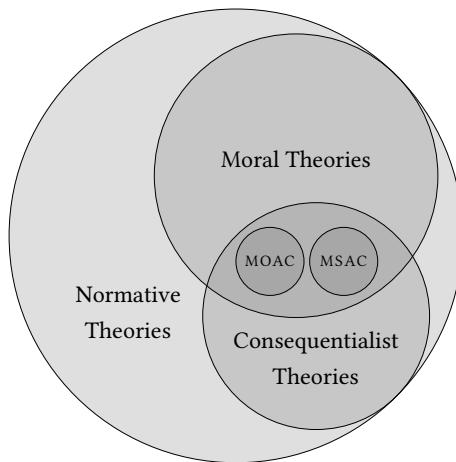


Figure 4.11: Rough-and-ready picture of the relation of normative, moral, and (different kinds of) consequentialist theories. (As an example of a non-moral consequentialist theory, the reader might think of classical theories of instrumental rationality.)

#### 4.3.4 A Counterexample Against DEONTIC COMPLETENESS?

I introduced the criteria in Table 4.4 without presenting many arguments for them. In doing so, I have mainly relied on their intuitive appeal in light of consequentialist background assumptions. How dangerous this can be shall

<b>Requirements</b>	CONCEPTUAL DEONTIC CONSISTENCY, METHODOLOGICAL INDIVIDUALISM, BEING A MOAC THEORY
<b>Desiderata</b>	RESOLVABILITY, NO MORAL DILEMMAS, WEAK DEONTIC COMPLETENESS, DEONTIC COMPLETENESS
<b>Comparative Cachets</b>	(CORE) EXTENSIONAL ADEQUACY, PARSIMONY, SIMPLICITY

**Table 4.4:** A non-final list of criteria I use below as a yardstick for assessing proposed solutions. Note that I add the weak *and* the strong versions because I consider a theory that has the strong version of those properties better: since every theory that has the stronger version also has the weaker one as a matter of logical or conceptual entailment, every theory that has the stronger properties fulfills two desiderata instead of only one.

be shown by the following example,<sup>96</sup> which shall at the same time justify why I have added the properties NO MORAL DILEMMAS and DEONTIC COMPLETENESS only to our list of desiderata rather than making them requirements. Entertain the following case:

**Case 4.1 (EVIL THORSTEN)** Thorsten finds himself with a spare Euro, a trivial amount he wouldn't even notice missing. A beggar approaches him, for whom that Euro would make a significant difference. However, Thorsten possesses a particular aversion to acting as consequentialism prescribes. So much so that the personal disutility he experiences from adhering to consequentialist principles would be twice as impactful as the beggar's potential benefit.

Certainly, this decision situation seems somewhat suspicious, and the involved form of self-reference certainly already sets some alarm bells ringing. Is the liar paradox right around the corner (cf. Bolander 2017)?

---

<sup>96</sup>The example is the result of a discussion with Thorsten Helfer, and it is partially based on his ideas.

At first, we might start our deliberation with the following ›normal form‹, where we assume 10 as the utility the beggar gets from the Euro and, accordingly, a disvalue of -20 for Thorsten for doing what is right:

		Giving is right	Keeping is right
Thorsten	Give Euro	-10	10
	Keep Euro	0	-20

What is the right thing to do for Thorsten if this were an adequate representation? We could apply the technique of conditionalization from Chapter 3. If it were right to give the Euro, then it would be better to keep it. Thus, keeping would be the right thing to do, and giving would be wrong – a contradiction. Thus, it cannot be that giving is the right thing to do. This should move us to the claim that keeping must be the right thing to do. But then we run into a similar contradiction: If it were right to keep the Euro, it would be better to give it. Thus, giving would be the right thing to do, and keeping would be wrong. Again, a contradiction. Thus, it can also not be right to keep the Euro.

This example might seem a little shady, and some might argue that it is ill-posed. True, self-references are always a red flag. But note that it seems to be a well-defined decision situation: There is an agent with two options, and even the consequences of these options are clearly specified. All that is problematic comes with the consequentialist framework: the qualities of the consequences are made to depend on what is the right thing to do according to consequentialism – which is a theory that, in turn, makes the rightness of an action a function of precisely these qualities. Thus, there is a recursion built in, which is not due to the decision situation's structure but to the con-

sequentialist theory's features. Insofar it seems as if the ›impossibility‹ of a proper description of the situation is due to the consequentialist and not the description of the case per se.

Consequentialists face two possibilities: either they must argue that cases like **EVILTHORSTEN** are ill-posed without them being ›at fault‹ (and I do not see how they could), or they need to go one of these three paths:

- Either they must live with the fact that at least one of the two options is right *and* wrong. In this case, they could argue that, for instance, giving is right and, therefore, keeping the Euro is better than giving it, such that giving it is wrong (and, hence, also not right). Doing so is to give up **CONCEPTUAL DEONTIC CONSISTENCY** (and, thus, also **CONSISTENCY**, at least if we stick to the (restricted) **CONSEQUENTIALIST STANDARD VIEW**).
- Or they must accept that both options are wrong. Then the problem with **EVILTHORSTEN** vanishes, as the correct way to think about the case would be captured by:

		giving and keeping are both wrong	
Thorsten	Give Euro	10	
	Keep Euro	0	

This solution comes with both the pain of rejecting **STRONG FREEDOM OF MORAL DILEMMAS** and the inconvenient question of how, then, it cannot be right to give the Euro. After all, this option had better

consequences, so according to MOCOR, it must be the right thing to do according to MOAC, which, again, would violate CONSISTENCY.

- Or, finally, they must accept that the options have no moral status *simpliciter*. Then the problem with EVILTHORSTEN vanishes for similar reasons as above, as the correct way to think about the case would be captured by:

		neither giving nor keeping is right
Thorsten	Give Euro	10
	Keep Euro	0

Again, they would need to explain why it is not giving that is the right thing to do in light of MOCOR. But they could do so by excluding EVILTHORSTEN from the domain of MOAC theories, implying giving up WEAK COMPLETENESS.

I am still unsure what lesson to learn from EVILTHORSTEN, but I think there are only two acceptable paths. Either camp MOAC finds a reasoned way to reject EVILTHORSTEN as being ill-posed, or they give up WEAK COMPLETENESS. All other options come with unbearable theoretical costs. As argued above, no serious theory should allow inconsistent assessments, and no serious theory should deny its own criterion of rightness based on *ex-ante* determinations. Because it's better to be safe than sorry, I suggest MOAC embrace DEONTIC COMPLETENESS as desiderata rather than as requirements. For the remainder of this project, however, I exclude cases involving self-references of moral theories.

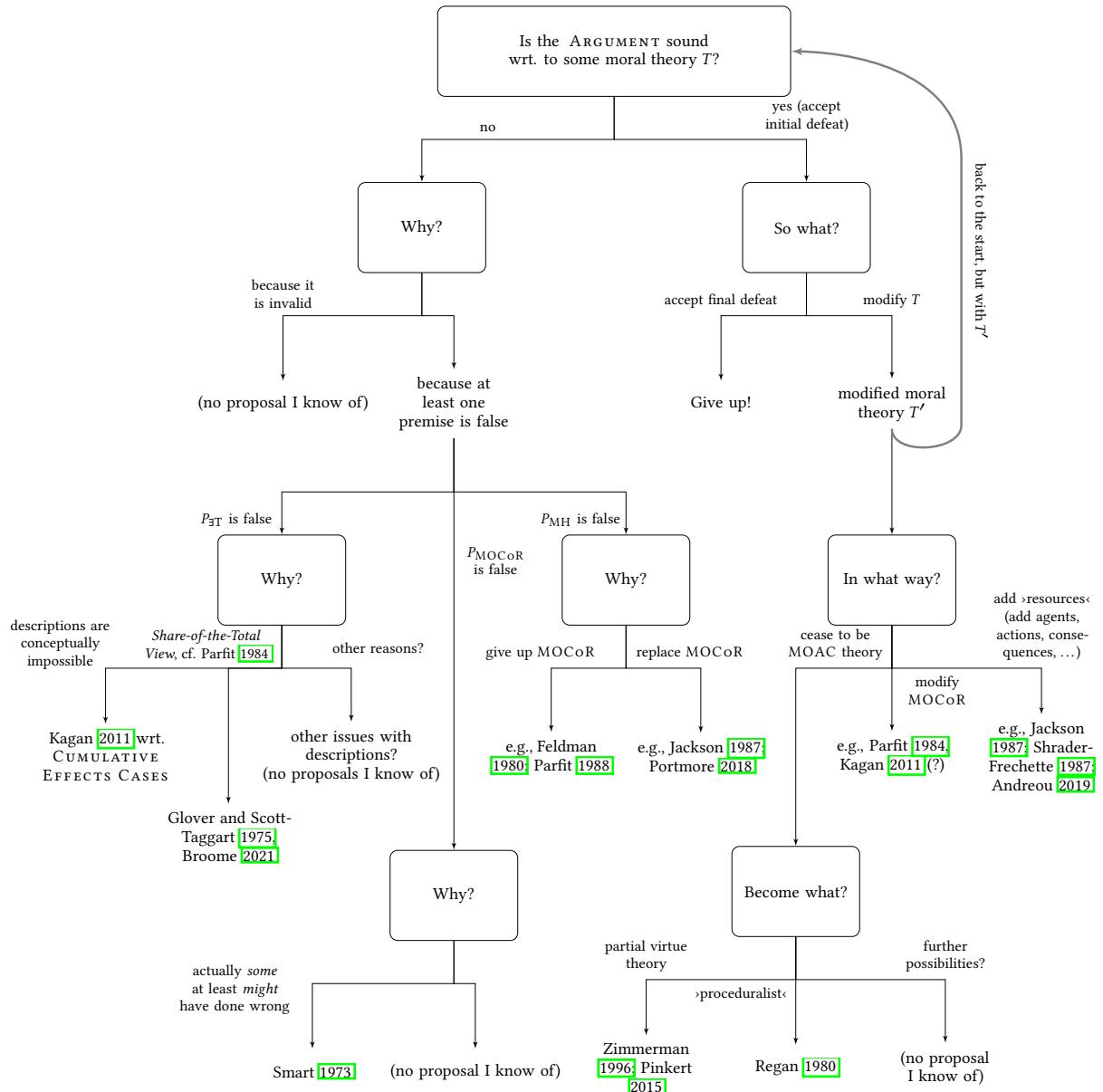
Before I develop my own approach to the CHALLENGE in the next part of this thesis, I collect some reasons why there is no convincing objective-consequentialist solution to the CHALLENGE so far.

#### 4.4 An Unsatisfying Exploration of the Solution Space

I contend that current proposed solutions to the CHALLENGE fall short of offering a comprehensive resolution. Nevertheless, I will *not even attempt* to substantiate this claim with a general argument. This section does *not* strive for exhaustive coverage of all potential solutions *nor* for an in-depth analysis of any approach. Any such endeavor would require entire chapters, if not a book. It would disproportionately extend the reconstructive nature of this work, limiting space for the development and unfolding of my own approach.

However, Figure 4.12 categorizes a carefully chosen set of approaches to the CHALLENGE<sub>int</sub>, the apex of the PYRAMID. I chose what I take to be either well-regarded or potentially overlooked, but particularly informative approaches. Notably, to the best of my knowledge, no one has critically questioned the soundness of the ARGUMENT by directly challenging its validity.

In the following, I highlight three handpicked approaches, though. Firstly, I touch upon Shelly Kagan's proposal (Kagan 2011), credited for rejuvenating recent discourse surrounding the CHALLENGE. However, we can set aside Kagan's perspective due to its early deviation from objective core tenets. Next, Felix Pinkert's particularly well-anchored contribution (Pinkert 2015) is briefly assessed and subsequently dismissed for straying too far from core act-consequentialist principles. Finally, I delve into an older, under-discussed (or often overlooked) approach by Frank Jackson (1987). Despite its apparent shortcomings, Jackson provides an relevant observation that bridges over to



**Figure 4.12:** The solution space for the CHALLENGE<sub>int</sub> (respectively, the ARGUMENT), annotated with a selection of (non-)solutions.

the second part of this project (where we will later, in addition, turn to J.C.C.

Smart's approach, cf. Smart 1973).

Through this limited yet targeted exploration, I aim to illustrate that the CHALLENGE remains unresolved, justifying the necessity of the subsequent segments of my thesis.

### 4.4.1 Kagan's Revived Discourse

Shelly Kagan's engagement with the CHALLENGE has undeniably rekindled interest in and the discussion of the CHALLENGE. Even though Kagan's approach is emphatically systematic, it is not apparent which variant of the CHALLENGE Kagan actually addresses, the CHALLENGE as the CHALLENGE<sub>int</sub> or the CHALLENGE as the No-DIFFERENCE CHALLENGE. First, Kagan suggests that he considers MOAC when he says (Kagan 2011, p. 107) that »the consequentialist is indeed concerned solely with the production of the best possible results«. We have already seen that this is not true for subjective variants of consequentialism (recall THE DRUG and MSAC). He then continues as follows (*Ibid.*, p. 107):

There is [a] kind of case that might reasonably be thought to be problematic even from the perspective of consequentialism. These cases appear to have the following structure: A certain number of people – perhaps a large number of people – have the ability to perform an act of a given kind. And if a large enough group of people do perform the act in question then the results will be bad overall. However – and this is the crucial point – in the relevant cases it seems that it makes no difference to the outcome what any given *individual* does. And this is true regardless of whether others are doing the act or not. Thus, if enough people do perform the act the results are bad overall; but for all that, it remains true of each individual agent that it makes no difference to the overall results whether or not *they* perform the act in question.

Several details of this presentation can be criticized as slightly misleading: for one, Kagan writes of »overall bad« and not of suboptimal results (even overall bad results can please the consequentialist, if they are the best possi-

ble ones); or that Kagan here surprisingly speaks of acts »of a given kind«, although it does not matter what kind of action is performed, only what kind of dependency results in what kind of valuative profile. But these minor inaccuracies should not distract us from the fact that Kagan definitely has a version of the CHALLENGE at hand. But which one?

I think that the most charitable reading is that Kagan is actually considering the NO-DIFFERENCE CHALLENGE from a *subjective* point of view – even though this is not absolutely clear.

First, consider a passage that apparently speaks against this reading (*ibid.*, p. 107).<sup>97</sup>

Intuitively, after all, in cases of the kind we are now turning to, the acts in question need to be condemned because of the results that eventuate from everyone's performing them.

This does not sound very subjectivist, even though it might be compatible with both, a focus on the CHALLENGE<sub>int</sub> and a focus on the NO-DIFFERENCE CHALLENGE.

However, next to this, we find a passage that even suggests that Kagan is more interested in a version of the CHALLENGE<sub>int</sub> (*ibid.*, p. 108):

The problem, in effect, is this: consequentialism condemns my act only when my act makes a difference. But in the kind of cases we are imagining, my act makes no difference, and so cannot be condemned by consequentialism – even though it remains true that when enough such acts are performed the results are bad. Thus consequentialism fails to condemn my act. In cases of this sort, therefore, consequentialism seems to fail even by its own lights.

---

<sup>97</sup>Again, it is misleading to frame the CHALLENGE as being connected somehow to the fact that there is an option of which it is true that »everyone's performing them«. Sufficiently many agents suffice.

Kagan's proposed *solution*, however, makes clear the subjective focus of his perspective (Kagan 2011, p. 119):

How, then, can the consequentialist condemn my act? The key to the answer lies in the thought that it is only overwhelmingly likely that my act made no difference. It is unlikely, but possible, that it did make a difference—that my own act was the triggering act. But if it was, then of course it made a very significant difference indeed, for the triggering act brought about the various bad results. What we have, then, is a familiar case of decision making under uncertainty. I cannot know for sure that my act brought about the bad results – indeed, I can know that most likely it did not: but even when I discount the overall bad results for the high likelihood that my act did not bring them about, the net result of doing this remains negative. That is, my act has a negative expected utility. And that is why, from a consequentialist perspective, it should not be done.

Kagan talks here about what one can *know* as an agent, what one has to assume, and that it is a decision under *uncertainty* that we encounter here. All these epistemic notions must be read as evidence that Kagan is *not* trying to make the methods of subjective consequentialism fruitful for MOAC, but is *actually* seeking to defend subjective consequentialism.

But then he cannot have the CHALLENGE<sub>int</sub> in his crosshairs, for we have long since seen that this version of the CHALLENGE is not at all pertinent to subjective varieties of consequentialism – quite simply because they do not accept MH, that is, because they accept the compatibility of acting rightly and suboptimal outcomes, at least in cases involving, for instance, incomplete knowledge (recall Jackson's THE DRUG example again). Thus, for the sake of the coherence of his work, Kagan must be considered as being concerned with the CHALLENGE as the NO-DIFFERENCE CHALLENGE.

This is not yet to say, of course, that we cannot use Kagan's approach for MOAC as well. But just one more thought shows us that we *cannot do so trivially*: where would we get the probabilities necessary to compute expectation values if they are not subjective probabilities? Would they be objective probabilities? How did these enter into an objective consequentialist framework and what commitments come with them? Kagan says nothing about these challenges – simply because he has no intention of defending objective consequentialism in the first place. But since that is the goal of this book, we can put this part of Kagan's approach to rest for now.<sup>98</sup>

#### 4.4.2 Pinkert: Modal Virtue Consequentialism

Unlike Kagan's account of the problem, we have discussed Pinkert's account already in some detail (Section 3.5.2.2). Therefore, there can be no question about which version of the CHALLENGE he has in mind. Pinkert has even proposed his own explication of PMH, called ON-THE-HOOK (cf. Section 3.5.2.2), which is meant to go beyond MOAC in its applicability. In this respect, it is clear that Pinkert wants to solve the CHALLENGE as CHALLENGE<sub>int</sub>. He makes his solution exemplarily explicit in the form of a new criterion of rightness<sup>99</sup> (Pinkert 2015, p. 982):

##### **Principle 4.4 (Modally Robust Act Consequentialism (MRAC))**

*An agent acts rightly if and only if the agent acts optimally in the actual world, and it should be such that for all possible combinations of the actions of other*

---

<sup>98</sup>My interpretation fits with Brian Hedden's who defended and extended Kagan's approach in a later piece (cf. Hedden 2020). Further, we come back to the question of how far one can actually apply expected utility-based considerations in the context of the CHALLENGE in the next part of this project.

<sup>99</sup>To be more precise, Felix Pinkert formulates a ›criterion of oughtness‹. His own formulation, therefore, differs slightly from the one presented here in starting: »An agent ought to act optimally in the actual world,...«.

*agents, if that combination were instantiated, they would act optimally in these circumstances.*

MRAC is an intriguing blend of act consequentialism with modal considerations (dealing with possibilities across different conceivable worlds). The first part, according to which an action is right if and only if, in the actual world, it leads to the best possible outcome, is basically MOCOR. The second part, which concerns ›modal robustness‹, introduces a rigorous requirement. Not only should the agent’s action be optimal in the real world, but it should also be such that in every conceivable scenario where other agents might act differently, the agent would still act optimally under these circumstances. This ensures that the agent’s action isn’t just accidentally doing right but is robustly optimal across various possible contexts.

In my opinion, Pinkert’s solution has one crucial flaw: although the condition looks like an objective-consequentialist one at first glance, on closer inspection, it is not. Recall the formally precise condition for objective-consequentialist theories:

**Definition 2.3 (Objective Consequentialist Theory (formal))** *T is an objective consequentialist theory if and only if it embraces an axiological sub-theory  $T_{Ax}$  with a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and an objective consequentialist criterion of rightness  $T_{COR}$  such that, for all decision situations  $D \in \mathcal{I}$  and for all  $\phi \in \Phi_D : D, C \models_T R\phi$  if and only if  $T_{COR}(\phi)$ .*

*A criterion of rightness  $T_{COR}$  is objective consequentialist if and only if, for all  $D \in \mathcal{I}$  with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  (with D’s actual context C)  $T_{COR}$  corresponds to a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :*

$$D, C \models_T R\phi \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))).$$

Thus, for a criterion of rightness to be genuinely objective consequentialist, it needs to be based *solely* on the valuative profile of the situation (i.e.,  $\text{Val}(\mathcal{O}_{D,C})$  and the value of the action under consideration (i.e.,  $\text{Val}(\text{Out}_C(\phi))$ ).

There cannot be such a predicate for Pinkert's MRAC. The modal robustness component makes it so that this predicate would also need to account for the optimality of actions *across* possible worlds. Assessing whether an agent acted rightly thus requires more than looking at the value of that action and the quality of the possible outcomes of the decision situation given the *actual* context. It also needs information about what action *would* have been taken instead if the other agents *had acted differently*. Thus, it needs a reference to a disposition (or something similar) of the acting agent, and thus, it needs another parameter, namely the agent of the action. This creates a kind of virtue-theoretical character (in the sense of the distinction in Section 2.3), and Pinkert's theory thus ceases to be a purely objective-consequentialist one – and therefore, in the sense of this project, is not a viable solution (Figure 4.12). In other words (and Regan's *lingo*), MRAC is not ›exclusively act-oriented‹.

Pinkert himself is well aware of this connection between the modal part of MRAC and virtuous character traits (as he tells us already in his abstract: »I interpret this Modally Robust Act Consequentialism as Act Consequentialism plus a requirement of moral virtue«, cf. Pinkert 2015, p. 971). My argument here is only meant to bring this to the point and, thus, to explicitly distinguish Pinkert's solution as unacceptable in terms of pure objective-consequentialist doctrine (and thus this project). I don't think Pinkert himself would have much of a problem with this. And even I would agree that, as long as there is no *purely* objective-consequentialist solution to the CHALLENGE, Pinkert's approach is probably the most acceptable for proponents of MOAC – even

if it would mean a move within the theoretical landscape. However, I hope to be able to present a solution in the second part that makes such a move completely unnecessary. At this point, it should only be stated that accepting Pinkert's solution would imply making a move.

#### 4.4.3 Jackson: Collectivism

Finally, we turn to Frank Jackson's take on the CHALLENGE. Jackson's position is essentially a direct response to Parfit's proposal in *Reasons and Persons* (Parfit 1984),<sup>100</sup> where Parfit suggested the following extension of the consequentialist framework (*ibid.*, p. 70):

(C7) Even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people.

Jackson believes this to be mistaken and brings up several quite convincing examples against (C7), which need not interest us here in detail. Ultimately, Jackson advocates (Jackson 1987, pp. 100-101) that the CHALLENGE is rooted in our »tunnel vision«, i.e., that we are »restricting ourselves, without fully realizing it, to the individual actions« while, instead, we should »enlarge the class of actions which may be morally evaluated to include group actions as well as individual actions« because then »we can say that the agents' group actions, though not their individual actions, are wrong.«

In sum, Jackson's position is this (*ibid.*, p. 101):

---

<sup>100</sup>Parfit's reaction, in turn, is directed primarily at Glover (1975) and even more so at Regan (1980). To Jackson, in turn, Parfit has reacted with an unpublished piece, also with an approach that goes more in the direction of ›biting the bullet‹ (Parfit 1988). See also the overview in Figure 4.12.

Parfit wants to enlarge our conception of what makes an action wrong with his guilt by association theory. I am suggesting that we respond to the difficult cases for the DIFFERENCE PRINCIPLE by enlarging our conception of what kinds of actions can be wrong (and right).

What may *prima facie* look like an elegant solution, comes with several weaknesses. First, Jackson's approach blatantly violates a requirement:

**Principle 2.1 (METHODOLOGICAL INDIVIDUALISM)**

*The primary bearers of deontic status are options of moral agents. Whatever has a deontic status and is not itself the act of a moral agent has this status merely in a derivative sense, that is, that deontic status is a function of the deontic status of certain options of moral agents.*

This theoretical well-founded principle seemed *prima facie* justified and a valuable piece in quite a lot of theoretical work. No action without agent – and, maybe even more importantly, no wrong-doing (or right-doing) without agent as well. Jackson *at least* owes us a good argument as to how a departure is justified and not just an *>ad-hoc dodge<*.

Second, it is not clear how Jackson's approach should help with the CHALLENGE as the PMH-based CHALLENGE<sub>int</sub>. Since PMH is based on the idea that consistently right action must be accompanied by the best results, the question must be asked how a wrong combination of actions, to which no agent corresponds, should help. In the end, even if we found a wrong *combination*, we would not have necessarily found a wrong *action*.

However, Jackson's approach might still help to master the CHALLENGE as the NO-DIFFERENCE CHALLENGE and in particular as the TRILEMMA as  $H_2$  would then simply be false, recall

( $H_2$ ) If something wrong *happens*, then because someone *did* wrong.

According to Jackson, any pre-theoretic acceptance of  $H_2$  would just be explained by our ›tunnel view‹. This seems not implausible to me. Still, it does not help with the CHALLENGE as the CHALLENGE<sub>int</sub>.

Third, the question arises of how the moral status of a combination of acts relates to those very acts' moral statuses. (Depending on Jackson's answer, this might even help him with the first point.) On this point, Jackson himself had quite a bit to say, and it is worthwhile, in view of the following part of the book, to have a short look into that.

Before I let his thoughts blossom, however, it should be summarized first that while Jackson's approach seems simple and charming, it comes with a lot of theoretical baggage. In a sense, similar things apply to Jackson's approach as to Pinkert's: The approach is not yet definitely out of the game. If there is no better solution, the work of spelling it out in detail might be worthwhile for proponents of MOAC. But the cost would be high, and the necessary justificatory work seems quite extensive to me. Furthermore, the likelihood of finding *new* challenges and counterexamples that might come with the new theoretical liabilities is probably high. So overall, it should be worth looking for an approach that comes with fewer commitments.

But before we use this impulse to move to the second part of my project, let us return briefly to the connection between the moral status of a combination of actions and the moral statuses of these individual actions: Jackson claims (Jackson [1987], p. 101) that »the moral standing of a group act can be partially or totally at variance with the standings of its constituents.« Here is an example Jackson uses to make his point (*ibid.*, p. 102):

**Case 4.2 (INTERSECTION)** *You and I approach an intersection from different directions. I have the right of way, so that what ought to happen is that you give way together with my driving straight on.*

The following normal form might be an appropriate representation of INTERSECTION as Jackson considered it:

		You	
		drive	stop
I	drive	worst	best
	stop	second-best	second-worst

INTERSECTION is indeed an asymmetric COORDINATION CASE.<sup>101</sup> Much more interesting is how Jackson *argues* for what is right and wrong in this case, given a certain combination of actions and what he thinks should be said about the moral status of that very combination. Here is Jackson's argument that, given that we both drive, the combination of [you stopping and me driving] is right, while my driving is wrong (*ibid.*, p. 102):

What in fact happens is that you do not give way, and would not regardless of what I do; so that were I to drive straight on, there would be an accident. What ought I to do? Drive straight on, consoling myself with the thought that I will be able to say from my hospital bed that I was in the right? Obviously, what I ought to do is stop. The position, then, is that the right group action is your

---

<sup>101</sup>The property of (a)symmetry will be discussed later in more detail. For now, the intuitive understanding is sufficient: a case is symmetric if we can exchange the agents and nothing changes for them with respect to the moral qualities of the consequences of their options. This is the case with Two FACTORIES, but not with INTERSECTION, because according to Jackson's construction, it is better if the one who has the right of way drives and the other one stops than the other way around. Thus, if you had the right of way, it would be better when you drive, and stop, and thus, the opposite of how things are in the original example.

stopping together with my driving on; the right action for you is to give way, and the right action for me is to stop. But if the right action for me is to stop, the wrong action for me is to drive on. Hence, we have a group action – your stopping together with my driving on – which is right, which nevertheless has a constituent action – my driving on – which is wrong.

It is to be anticipated that some readers will want to object to Jackson's argument along these lines: of course, it's right for me to stop, but that's *because* you're not doing what you're supposed to do. And yes, in a certain sense, it is correct that it would be right for you to stop and for me to drive (because I have the right of way, or for whatever other reason), but this *presupposes* that this combination is still a choice at all (for whom, actually?). As soon as it is set that you drive on, this combination is no longer on the table. What this kind of objection shows, I think, is that we have to distinguish carefully between the *initial state*, in which it is not yet set what somebody does, and what is the case *as soon as somebody has done something*. We need *space for dynamics* – and the consequentialist does not have that conceptual space in his understanding of COORDINATION CASES so far. Since camp MOAC hasn't that conceptual resource (yet!), Jackson can counter this anticipated objection with the following straightforward move (Jackson [1987], p. 102):

I have been surprised by how often I have met the following response to this sort of example. »The argument turns crucially on the claim that I ought to stop. But I ought to stop *only because* you do not do as you ought, namely, give way, and so all that is really true is that I ought to stop given you do not give way.« However, the sketched alternative position is inconsistent: »*P* only because *Q*« entails *P*! To grant that it is true that I ought to stop is true only because you do not give way is *ipso facto* to grant that it is true that I ought to stop.

Of course, Jackson is right that  $p$  because  $q$  implies that  $q$ .<sup>102</sup> Only I don't think this really helps him substantiate his claims. But to get to the heart of why Jackson's argument doesn't hold, we need a *richer* understanding of COORDINATION CASES and collective decisions in general, an understanding that allows us to think about sequences of actions, to truly decompose TROUBLEMAKERS into sets of individual decision situations, and to make room for the assessment of combinations of actions. The journey to this deeper understanding comes with some surprises, also with respect to the CHALLENGE and its validity, but it shows the sustainable way out of the misery of consequentialism in multi-agent settings. We embark on this journey in the second part of this thesis.

---

<sup>102</sup>For a well-developed logic of because see e.g. Schnieder 2011.



## **Part II**

# **The REAL CHALLENGE (and How to Solve It)**



## Overview of Part 2

In the second part of this book, I develop my very own approach to the CHALLENGE. I will do so in two steps. The first step reveals a serious misconception underlying the CHALLENGE as reconstructed in the first part of my project. Thus, I suggest understanding the CHALLENGE rather as a symptom of a deeper cause, i.e., as a consequence of an implicit move consequentialists have made in order to avoid another challenge that I will call the REAL CHALLENGE. The second step, then, is to offer a solution to the REAL CHALLENGE that does *not* result in the CHALLENGE. Proceeding in this way, I master several challenges for consequentialism at once: I demystify and deconstruct the well-known and much-discussed CHALLENGE and, at the same time, solve the deeper-lying and so far only sporadically discussed REAL CHALLENGE. In the end, both challenges are off the table and MULTI-AGENT CONSEQUENTIALISM is born.

After extending the already introduced formalism to the domain of collective decision situations in Chapter 5, I argue in Chapter 6 that something is seriously wrong with the CHALLENGE by revealing that the ARGUMENT involves an equivocation and, ultimately, should be considered invalid. Roughly speaking, it is questionable whether the right actions derived in  $P_{MOCoR}$  really stand in conflict with the missing wrong action required in  $P_{MH}$ . A closer

look at the two premises raises the question of *whether the same context of evaluation* is relevant within the two premises.  $P_{\text{MOCOR}}$  explicitly assesses retrospectively, i.e., *ex-post*: given what the others did (or will do, ..., cf. Section 3.5.2.3), none of the agents could have changed anything for the better. For this reason, each of them has acted rightly, given the actions of the other agents, but only in retrospect.  $P_{\text{MH}}$ , however, apparently results from an aspiration to ›guide‹ behavior, namely, in the direction of the best outcome. I thus argue, thus, that the best reconstruction of the ARGUMENT is invalid. I close the chapter with the somewhat surprising conclusion that this result is not as good for MOAC advocates as one might initially think. After all, this result is a Pyrrhic victory, revealing what I call the REAL CHALLENGE. This challenge consists, put somewhat roughly, in the diagnoses that MOAC does not give incorrect assessments but no (unconditional) assessments at all. I argue that the REAL CHALLENGE is even worse than the original CHALLENGE as it means that i) MOAC is seriously deontic incomplete and ii) MOAC theories, at the end of the day, still ultimately violate the spirit of PMH: if there are no genuinely right actions, we cannot hope that morality ›highlights‹ or points toward the way to the best possible outcome. Revisiting Regan's ›impossibility result‹, I conclude the chapter by explaining why we should understand the CHALLENGE as a symptom of the consequentialists' approach to fill the deontic gaps in certain collective decision situations. In the remainder of this part, what remains is to find a way for camp MOAC to fill the gaps *without* invoking the CHALLENGE.

In Chapter 7, I suggest that consequentialists have overlooked (or forgotten about) a plausible candidate for intermediate outcomes in collective decision situations, namely the decision situations of other agents. Following

this APPROACH brings new material to the consequentialist workbench. I show that his fresh perspective allows a unified understanding and representation of COORDINATION CASES and SEQUENTIAL CASES in what I call generative extensive forms. From all this, it follows that camp MOAC now just has to decide *how* they want to fill the deontic gaps and, thus, overcome the REAL CHALLENGE, hopefully without stumbling into the CHALLENGE again. I introduce some possible theoretical puzzle pieces that might allow camp MOAC to do so that I call *collective amendments*. However, since we lack a clear idea of how to assess them, we cannot, at this point, decide between them.

To overcome this last remaining challenge, I revisit the PMH in Chapter 8. Based on yet another kind of COORDINATION CASE, I argue that the current explication, MH, must be relaxed because a theory adhering to MH would necessarily conflict with an even more basal and persuasive principle, i.e., the principle of NORMATIVE SUPERVENIENCE, a principle logically entailed by the very definition of objective consequentialism. This opens a new perspective that allows us to explicate the overarching goal consequentialists should aim for when filling the formerly diagnosed deontic gaps. Based on the idea of calculating the expected value of the adoption of certain amendments, I propose a formal framework for amendment ranking built upon the idea of policies from formal decision theory. After defining a ›testbed‹ of important *structural types* of collective decision situations, I perform some calculations to determine a winner.

Finally, I conclude my project in Chapter 9 and point at some potential directions for related future research. I end by reproaching the deontologists that they are the ones struggling with collective contexts now and invite

them to at least move one foot into camp MOAC to benefit from the new conceptual possibilities that have been developed in my work.

Similarly, as in the case of the first part, it makes sense to send some formalities in advance to avoid introducing new concepts along the argumentation drop by drop, which would seriously disturb the flow. The next chapter, thus, essentially introduces an extension of decision situations to collective contexts. These at least semi-formal definitions and concepts will play a more central role in this part than formalism has done in the first part.

# Chapter 5

## Preliminaries II

At this point, we have encountered two types of collective decision situations, **SEQUENTIAL CASES** and **COORDINATION CASES** (cf. Figure 5.1 and Figure 5.2 for the corresponding kinds of **TROUBLEMAKERS**), and many specific instances of these types. So far, quite in line with the debate on the **CHALLENGE**, I have assumed that we can simply apply MOAC to the agents' decisions in such collective situations, a view I shall call **COMPOSITIONALISM**. As already marked earlier, behind this practice is the implicit assumption that, in a certain sense that deserves to be dragged to light, collective decision situations can somehow be decomposed into individual decision situations (or, at least, that one can identify individual decision situations for each agent).

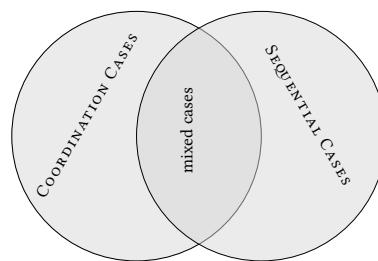


Figure 5.1: Kinds of collective decision situations.

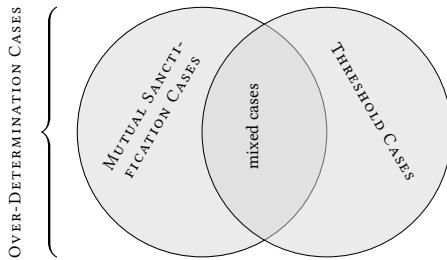


Figure 5.2: The set of all **TROUBLEMAKERS** as tackled by my project (see Section 4.2). As explained earlier, it suffices to focus on pure cases. Recall that **MUTUAL SANCTIFICATION CASES** are **COORDINATION CASES** while **THRESHOLD CASES** are **SEQUENTIAL CASES**.

On closer inspection, however, things are not quite so obvious, as we are not guaranteed to find ›proper‹, i.e., unconditional consequences for agents in collective decision situations. It is far from being obvious, thus, whether COMPOSITIONALISM is true or whether we should not rather adopt the GENUINE KIND VIEW, according to which at least some collective decision situations are decision situations of a genuinely own kind. Then, however, it remains far from being obvious what MOAC actually has to say in such situations. This question, hence, will turn out to be crucial for overcoming the CHALLENGE. Accordingly, this part begins with an in-depth examination of this very relationship of collective and individual decision situations.

To get a better understanding of this question and of the arguments encountered so far, I will again introduce some concepts and (semi-)formal notions. These allow me to concisely explicate certain properties of collective decisions, most importantly the TRIAD of *maximality*, *order-invariance*, and *symmetry*. These properties allow me, first, to precisely restrict my project in this regard to an actually manageable scope. Second, these properties later enable a systematic treatment.

Finally, I will investigate what MOAC (and, more specifically, MOCoR) can actually deliver with respect to COORDINATION CASES, introducing the notion of separability of collective decision situations along the way.

## 5.1 Collective Decision Situations

While normative and moral questions primarily concern individual decision situations, the CHALLENGE is raised by *collective* decision situations. Up to now, I have pretended (following the existing discourse) that trivially, each of

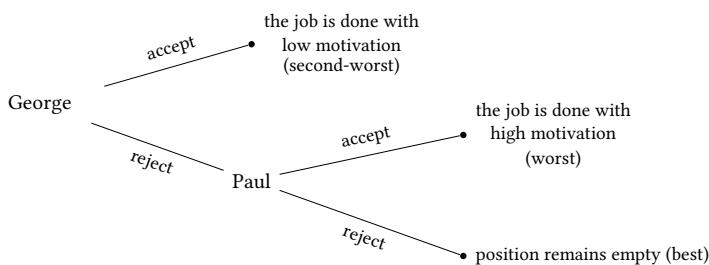
the agents involved in a collective decision situation faces his very own individual decision situation. We will see that this is, indeed, all but an innocent assumption.

We have already encountered several collective decision situations in this book, and we will meet several more. Thus, we have already developed an intuitive understanding of collective decision situations that we can now develop toward a more precise definition. Recall the rough-and-ready definition from Chapter 3:

### **Definition 3.1 (Collective Decision Situation (tentative))**

A collective decision situation is a situation in which multiple agents are each presented with multiple options, and within a given context, each combination of actions has an associated consequence.

In light of the situations we have encountered up to this point, this definition deserves some slight modifications. Most importantly, we should make conceptual space for the possibility that in some cases, agents only have to make decisions if other agents have decided or acted in certain ways. Recall, for example, JOB MARKET which was captured in extensive form like this:



It seems appropriate to say that in this **SEQUENTIAL CASES**, there are two agents, George and Paul, but only George is *guaranteed* to be put into a situation where he has to decide. Paul may have no choice to make at all because if George takes the job, it won't be offered to Paul. Thus, the combination

of George accepting the job and Paul likewise accepting it is not a *possible* combination of actions according to JOB MARKET. However, both agents are needed for a sufficiently complete description of the situation with respect to the possible, relevant consequences. For example, both the action combination of George refusing the job and Paul accepting it and the ›combination‹ consisting of merely one action, namely George accepting the job, are associated with genuine outcomes.

Let it be agreed that a combination of actions is called *proper* (relative to a concrete collective decision situation) if there is a specified outcome for it. Further, I will describe choices like Paul's such that an agent is *potentially* presented with multiple options. We can then define collective decision situations a bit more precisely like this:

**Definition 5.1 (Collective Decision Situation)** A collective decision situation is a situation in which multiple agents are each (potentially) presented with multiple options, and within a given context, each proper combination of actions has an associated consequence.

All this is *not* to exclude that we can hypothesize about purely hypothetical actions, i.e., actions that agents would have taken if they had the chance. Thus, hypothetical actions can very well be relevant for moral assessments and other considerations. As we will see, to understand all this – potential choices, proper combinations, moral assessments based on and of hypothetical actions, ... – is crucial for consequentialists in order to be able to master the CHALLENGE. Achieving this understanding, however, is a non-trivial endeavor. In the remainder of this section, I prepare the ground that will serve us in the remainder of this part to attain that understanding.

A note at the outset: for the sake of readability and comprehensibility of many of the following considerations, I will not attempt to address the CHALLENGE in its greatest generality. Instead, I will make some simplifications and limitations. I am sure, even after repeated review of my results, that generalization is a mere technicality and formal exercise. In the end, this is a dissertation in analytic philosophy that is visibly unafraid of formalisms. It is not, however, a formal treatise in mathematical multi-agent ethics.

The first limitation of this kind has already been indicated earlier. I mainly concentrate on the ›smallest‹ collective decision situations, i.e., situations with two agents and two possible actions each. This makes it easy to represent these cases, and everything I have to say is transferable to cases with an arbitrary number of agents with an arbitrary number of options. This is so because we can generalize findings to such ›larger‹ cases through rather straightforward inductive reasoning, as soon as we have introduced sufficient suitable structure. I explain this in a bit more detail when I sum up my approach at the end of this book. However, this presupposes a formal structure on which one can operate mathematically accordingly.

Before I turn to introducing such structure through formal specifications and shorthands similar to those introduced earlier in the context of individual situations, we should pause for a moment and ask whether we *really* need to introduce another kind of decision situation. In other words, aren't individual decision situations sufficient, possibly just completed with some structural information about dependencies between agents, actions, and consequences?

Two questions are in order, only one of which can be immediately shelved. First, whether there are or can be other types of decision situations besides individual and collective ones. This is to ask whether there are possibly even

more kinds of decision situations. This question, I think, can be answered immediately in the negative, for lack of candidates.<sup>103</sup> Hence, for the remainder of this project, I operate under the assumption that there is no other third kind of decision situations.<sup>104</sup>

The second question is whether collective decision situations can be reduced to individual ones, i.e., whether we *really* need to consider a second kind of decision situation. This question is much more exciting and will occupy us more than once during this project. Two incompatible claims are the two most interesting possible answers to this question. First, consider

**Claim 5.1 (COMPOSITIONALISM)** *All collective decision situations can be reduced to individual decision situations (plus some structure).*

Note that COMPOSITIONALISM allows that collective decision situations are *more than just* individual decision situations. Even if reducible to individual decision situations, many collective decision situations have to involve *some* structural element, for instance, the temporal order and counterfactual dependencies between earlier and later decision situations. Nevertheless, according to COMPOSITIONALISM, a collective decision situation of  $n$  agents involves (at least)  $n + 1$  decision situations: the collective decision situation

---

<sup>103</sup>There is another distinction that may come to mind in this context: the distinction between decisions made by diachronic persons and those made by time slices (or temporal parts) of such persons. Probably the best-known case is that of Professor Procrastinate (cf. Jackson and Pargetter 1986; C. Woodard 2009; Jackson 2014) and the related actualism–possibilism debate in normative ethics (cf. Timmerman and Y. Cohen 2020). This distinction is either not at all different from the distinction between individual and collective decision situations, namely when time slices (or temporal parts) qualify as agents (cf. Dietz 2020), or it is orthogonal to that distinction. In either case, however, similar questions arise. For example, whether decisions of diachronic persons can be reduced to decisions of time slices. Because of this structural similarity to the CHALLENGE, I will briefly return to this debate at the end of this thesis.

<sup>104</sup>As with the different kinds of collective decision situations, there could, of course, also be mixed cases here, i.e., those which are clearly individual decision situations and possibly genuine collective decision situations. But as with other mixed cases, it is enough for us to understand the two pure cases to be able to handle mixed cases well.

and at least one individual decision situation for each and every agent.<sup>105</sup>

Compare COMPOSITIONALISM to

**Claim 5.2 (GENUINE KIND VIEW)** *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

GENUINE KIND VIEW is the negation of COMPOSITIONALISM. Since they both express clearly meaningful propositions (for example, they are not category mistakes), either one or the other is true. According to the GENUINE KIND VIEW, at least some collective decision situations cannot be reduced to decision situations completely. Thus, at least some agents in some collective decision situations are, then, not also in their ›own‹ individual decision situation but are agents only within some truly genuine collective decision situations.

The straightforward application of moral theories, which are defined over individual decision situations,<sup>106</sup> is only possible if COMPOSITIONALISM holds. Given COMPOSITIONALISM, we could, in principle, decompose arbitrary collective decision situations to the individual agent's decision situations, allowing us to apply moral theories to these situations to determine the deontic status of the agent's options. Trivially, this procedure would also align with METHODOLOGICAL INDIVIDUALISM (cf. Principle 2.1). Recall

### **Principle 2.1 (METHODOLOGICAL INDIVIDUALISM)**

*The primary bearers of deontic status are options of moral agents. Whatever has*

---

<sup>105</sup>There might be intermediate or partial collective decision situations involved, for instance, one or several situations of  $n - 1$  agents,  $n - 2$  agents and so on.

<sup>106</sup>It may be objected that I have formulated this definition in Section 2.2 accordingly, but that this does not necessarily correspond to the general understanding (if there is such thing) of moral theories. However, I claim that my definition actually corresponds largely to our everyday use of language and the practice in normative ethics (see also METHODOLOGICAL INDIVIDUALISM).

*a deontic status and is not itself the act of a moral agent has this status merely in a derivative sense, that is, that deontic status is a function of the deontic status of certain options of moral agents.*

On the contrary, if the GENUINE KIND VIEW were correct, there would be no guarantee that for any arbitrary collective decision situation, there is an individual decision situation for each and every agent. It is then far from obvious how we should apply moral theories to such inseparable collective decision situations, let alone in a way that is in line with METHODOLOGICAL INDIVIDUALISM (we remind ourselves of Jackson's approach, cf. Section 4.4.3). Thus, whether COMPOSITIONALISM or the GENUINE KIND VIEW is true is highly relevant for this project.

On closer inspection, however, COMPOSITIONALISM could turn out to be a hopeless position. We have already seen some collective decision situations where at least some arguably morally relevant parts of certain consequences are *not* fully determined by any individual agent's actions. For instance, with respect to our running example Two FACTORIES, even though both Ann and Ben can ensure the pollution of the river (by polluting individually), *none* of them can ensure that it is *not* polluted. This depends on what *both* agents do. This is to say that apparently some (parts of) consequences are defined *only* over *combinations* of actions. But if we cannot properly take into account these consequences in our moral assessments, we could, and in many cases, no doubt, actually do fail to acknowledge certain morally relevant consequences of actions. We thus apparently cannot reduce arbitrary collective decision situations to individual ones.<sup>107</sup>

---

<sup>107</sup>Some may argue that even such collective decision situations are reducible to individual decision situations, but to situations in which we cannot assign *all* consequences of combinations of actions to individual actions. However, such reductions or decompositions would

Although in the course of this book, I will try to defend COMPOSITIONALISM and, building on it, develop a powerful solution to the CHALLENGE, I will first put forward more evidence against COMPOSITIONALISM. Then, in the next chapter, I will base my fundamental criticism on these grounds. Hence, one of the goals of the following sections is to explicate a couple of relevant properties of collective decision situations and related distinctions, some of which pull COMPOSITIONALISM's defensibility seriously into doubt. Most importantly, we will try to get a precise understanding of the apparent non-separability of some collective decision situations, of which TROUBLEMAKERS are a subset by definition.

## 5.2 (Semi) Formalism and Shorthands

As before, with individual decision situations, it will prove useful to have some shorthands, formal specifications, and notions for collective decision situations and their components later on. Let there be some collective decision situation  $D$  of agents  $A_1, \dots, A_n$  with corresponding option spaces  $\Phi_{A_1}, \dots, \Phi_{A_n}$ . Let  $\mathcal{A}$  refer to the set of agents in  $D$  and let  $\Gamma$  be the set of these agents' option spaces, i.e.,  $\Gamma := \{\Phi_A \mid A \in \mathcal{A}\}$ . We use  $\Upsilon$  to refer to *proper* combinations of actions, possibly with superscripts, i.e., we write  $\Upsilon^1, \Upsilon^2, \dots$ , etc. or simply  $\Upsilon, \Upsilon', \dots$ , etc. to distinguish between different proper combinations. We model a combination of the actions  $\phi_1, \dots, \phi_n$  as an  $n$ -tuple  $\langle \phi_1, \dots, \phi_n \rangle$ . It is thus natural to use the index notation  $\Upsilon_i$  to access the  $i$ th element of a combination  $\Upsilon$ , i.e., if  $\Upsilon = \langle \phi_1, \dots, \phi_n \rangle$ , then  $\Upsilon_i = \phi_i$ . For convenience, we use the set-theoretic notation of  $\phi \in \Upsilon$  to express that there is an index  $i$  such that  $\phi = \Upsilon_i$ .

---

not be without *loss*. But since, from the point of view of at least consequentialist theories, essentially important aspects of the collective decision situation would be lost in such →*lossy reductions*←, this position is unacceptable from the outset, at least for champions of MOAC. This project is only interested in the general possibility of *lossless* reductions.

Next, we turn to the set of proper combinations of  $D$ . This set, which is clearly built on top of  $\Gamma$  and that I thus refer to by  $\Psi_\Gamma$ , I will call the *domain of  $D$* . It turns out that defining  $\Psi$  on a general level is rather complicated because not every combination of actions we can, in principle, construct from an option space is necessarily a proper one, i.e., a combination over which consequences are specified (remember **JOB MARKET** above). I will come back to some details in a short while.

As always, I sometimes leave out unnecessary indexes, e.g., I write  $\Psi$  for the domain of some collective decision situation if the relationship to the specific option space doesn't matter or is trivial. As with individual decision situations, we use  $C$  to refer to the set of relevant contexts of  $D$  and call  $\mathcal{O} = \{O_1, \dots, O_m\}$  the set of consequences of  $D$  (with  $m = |C| \cdot |\Psi|$ ). Recall that the set of relevant consequences is implied by the relevance stance under consideration (cf. Section 2.3.1). In the case of MOAC, we are thus dealing, in line with the **OBJECTIVE VIEW**, with exactly *one* relevant context, namely the actual one.

To be able to write succinctly about the relationship between combinations and contexts with the consequences in  $\mathcal{O}$ , we model this relation in terms of an *outcome function*  $\text{Out} : \Psi \times C \rightarrow \mathcal{O}$ . Sometimes, when more than one decision situation is under consideration, I use indices for disambiguation; where not needed, I leave them out.

As with individual decision situations, we use tuples as abstract representations of collective decision situations: a collective decision  $D$  for a set of agents  $\mathcal{A}$  with a set of option spaces  $\Gamma$  and an outcome function  $\text{Out} : \Psi_\Gamma \times C \rightarrow \mathcal{O}$  is represented by a tuple  $D := \langle \mathcal{A}, \Gamma, \text{Out} : \Psi_\Gamma \times C \rightarrow \mathcal{O} \rangle$  (or, for singletons  $C = \{C\}$ , just  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$ ). Again, just like

with individual decision situations, we often can refer to a collective decision situation  $D$  with a set of relevant contexts  $C$  and a domain  $\Psi_D$  just by giving one such tuple and, thereby, implicitly introducing both (and also the set of outcomes  $\mathcal{O}$ ) through explicating just the signature of the outcome function. Finally, let  $\mathbb{C}$  denote the set of all collective decision situations.

Valuation functions and related concepts can be straightforwardly extended to collective decision situations. Let  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  be a function that assigns arbitrary outcomes from the now extended

$$\mathcal{W} := \bigcup_{D \in \mathbb{C}} \mathcal{O}_D \cup \bigcup_{D \in I} \mathcal{O}_D$$

a value from some value space  $\mathcal{V}$ . For a set of outcomes  $\mathcal{O}_D \subseteq \mathcal{W}$  of some collective decision situation  $D$ , let us call  $\text{Val}(\mathcal{O}_D) := \{ \text{Val}(O) \in \mathcal{V} \mid O \in \mathcal{O}_D \}$  the *valuative profile* (of  $D$ ). Again, we require the existence of a total order  $\leq_{\mathcal{O}_D}$  over  $\text{Val}(\mathcal{O}_D)$  for arbitrary  $D \in (\mathbb{C} \cup I)$ . Finally, given an individual decision situation  $D$  such that  $\Upsilon \in \Psi_D$ , and a relevant context  $C$ , let  $\text{Val}_C(\Upsilon)$  be agreed upon as an abbreviated notation for  $\text{Val}(\text{Out}_C(\Upsilon))$ , exactly as it has been defined for the individual case.

Before we continue, let's quickly introduce a more sophisticated version of COLLECTIVELY MAXIMIZING, much as we did for its individualistic counterpart MOCOR in Part I. Recall

### **Principle 1.1 (COLLECTIVELY MAXIMIZING (tentative))**

*If all agents act rightly, then they are guaranteed to produce the morally best outcome they could bring about together.*

Based on the notions established in this section, we can clarify:

**Principle 5.1 (COLLECTIVELY MAXIMIZING)** *Let  $D$  be a collective decision situation with domain  $\Psi$  and with actual context  $C$ . If  $\Upsilon \in \Psi$  consists only*

*of right actions, then there is no (and cannot be) alternative  $\Upsilon' \in \Psi$  with better consequences than  $\Upsilon$  relative to  $C$ .*

It is worth digging a bit deeper with respect to proper actions and the domain, even though, for the most part, the combinations within the domain of some collective decision situation, i.e., the proper combinations relative to that situation, should be evident given a sufficiently complete case description. Nevertheless, having a more precise notion of the domain allows for pinning down specific properties that mark different classes of collective decision situations. These classes, in turn, help to make precise restrictions with respect to what kind of collective decision situations I focus on in this project (and *why* I do so). While making these restrictions increases this part's overall readability and comprehensibility, it admittedly comes with a loss of generality. As indicated above, however, this generality can be regained through formal efforts and hard but rather dull, purely technical work that I leave to those who love such undertakings or want to earn some diligence points.

### 5.2.1 Domains and Properties and the Triad

First, let me stress that we ignore the possibility of synchronous actions for now. I will return to this issue later in this part, but for now, we can lay that possibility, which seems rather esoteric anyway, aside for the time being as it would just complicate matters. In light of this restriction, we start by defining the most general domain possible. This *maximal possible set of combinations* given some set of option spaces  $\Gamma$  I shall refer to as *domain space*  $\Psi_\Gamma^*$ . All actual domains of given collective decision situations will necessarily be subsets of that set, i.e.,  $\Psi_\Gamma \subseteq \Psi_\Gamma^*$ . Assuming that the empty set is *not* a meaningful

combination – after all, it must be possible for choices to be made –, we may define that superset of domains as

$$\Psi_{\Gamma}^* = \bigcup_{\tau \in S_n} \bigcup_{i=1}^n \bigtimes_{j=1}^i \Phi_{A_{\tau(j)}}$$

This definition borrows the concept of *permutations* from combinatorics: Let  $S_n$  denote the set of permutations over the (index) set  $I_n = \{1, \dots, n\}$ , i.e., all bijections  $\tau : I_n \rightarrow I_n$ . A permutation  $\tau \in S_n$  can thus be understood as a rearrangement of the elements of  $I_n$ ; hence,  $S_n$  is the set of all possible rearrangements over that  $I_n$ .  $S_n$  is also known as the *symmetric group* of  $I_n$ .

Collective decision situations with a full domain, i.e., with  $\Psi_{\Gamma} = \Psi_{\Gamma}^*$  (given their set of option spaces  $\Gamma$ ), are not uncommon (think of **TWO FACTORIES**), but also far from being guaranteed, especially in the context of **SEQUENTIAL CASES** (recall **JOB MARKET**). Thus, there still remain a lot of different considerations concerning  $\Psi_{\Gamma}$ .

I'll put **SEQUENTIAL CASES** aside for now and limit myself to **COORDINATION CASES** before returning to **SEQUENTIAL CASES** at the end of this chapter. For **COORDINATION CASES**, I am going to focus, for the most part, on **TWO FACTORIES**-like cases that exemplify a couple of properties, most of which directly translate to their domains, which I will assume for large parts of this part of my project.

First, while these cases are *minimal* in that they involve two agents with two options each, they are *maximal* in that their domain only contains *maximal* combinations, i.e., combinations that contain one action for each agent.

Second, the moral quality of the consequences of these combinations is *order invariant*, i.e., not only is it possible for the agents to act in arbitrary orders, but the outcomes are, in a sense, morally indistinguishable. While this

is, strictly speaking, a property concerning the valuative profile of the decision situations more than of the domain itself, it requires all the corresponding combinations of actions to be proper in the first place.

Third, I focus on *symmetric COORDINATION CASES*, i.e., in a very rough sense, cases where the quality of the outcome depends only on the performed actions and not on *who* is performing them. All three of these properties will be specified and explicated in more detail below.

I call these three properties together the **TRIAD** and will say that a collective action case with all three properties satisfies the **TRIAD**. With a few exceptions explicitly marked as such, the investigations carried out in this part are limited to **COORDINATION CASES** that satisfy the **TRIAD**. This subclass of **COORDINATION CASES** is well-populated and contains all relevant **MUTUAL SANCTIFICATION CASES** typically discussed in the literature. However, this restriction naturally entails a loss of generality but, at the same time, a gain in readability and comprehensibility. However, the cases singled out in this way make excellent base cases for induction-based generalizations – as we will see also with respect to generalizing toward **SEQUENTIAL CASES**.

### 5.2.1.1 Maximality

Maximality is a straightforward property. We can capture maximality formally:

**Property 5.1 (Maximality)** *Let  $D$  be a collective decision situation with a set of agents  $\mathcal{A}$  with the corresponding option spaces  $\Phi_A$  for each  $A \in \mathcal{A}$ . A combination of actions is maximal (with respect to  $\mathcal{A}$ ) if and only if it contains a combination of each agent (from  $\mathcal{A}$ ).*

$D$  is maximal if and only if, in every proper combination of actions is maximal, i.e.

$$\forall \Upsilon \in \Psi_{\Phi_A} : \forall A \in \mathcal{A} : \exists 1 \leq i \leq |\Upsilon| : \Upsilon_i \in \Phi_A$$

For simplicity's sake, let us assume that no agent can perform more than one action, which is rather plausible in most cases. Against the background of this assumption, we can simplify the formal condition: A collective decision situation is maximal if and only if

$$\forall \Upsilon \in \Psi : |\Upsilon| = |\mathcal{A}|$$

In terms of the domain, maximality comes down to restricting the domain such that we get (for  $\Gamma = \{\Phi_{A_1}, \dots, \Phi_{A_n}\}$ ):

$$\Psi_{\Gamma, \text{maximal}}^* \subseteq \bigcup_{\tau \in S_n} \bigtimes_{i=1}^n \Phi_{A_{\tau(i)}}.$$

### 5.2.1.2 Order-Invariance

Two FACTORIES exemplifies another property: It is *invariant under the order of actions* (or, a little shorter, order-invariant), i.e., it does only matter *what* actions the involved agents perform but not *in which order* they do so. As with maximality, we shall capture order-invariance formally. For this, let us first define the set of proper combinations with the same constitutive actions as a given combination  $\Upsilon$  (relative to specific collective decision situations with domain  $\Psi$ , of course) – the *set of sequential recombinations* of a given combination – as follows:

$$\widehat{\Psi}_{\Upsilon} := \{ \Upsilon' \in \Psi \mid \forall \phi \in \Upsilon : \phi \in \Upsilon' \text{ and } \forall \phi \in \Upsilon' : \phi \in \Upsilon \}$$

Based on this notion<sup>[108]</sup> we can now define

---

<sup>[108]</sup>Note that  $\Upsilon \in \widehat{\Psi}_{\Upsilon}$  by construction.

**Property 5.2 (Order-Invariance)** *Let  $D$  be a collective decision situation with domain  $\Psi$ .  $D$  is invariant under the order of action (or, shorter, order-invariant) if and only if for every proper combination of actions, all sequential recombinations are proper, and the outcomes of all these combinations are valuated equivalent, i.e.,*

$$\forall \Upsilon \in \Psi : \widehat{\Psi}_\Upsilon \subseteq \Psi \text{ and } \forall \Upsilon' \in \widehat{\Psi}_\Upsilon : \text{Val}(\text{Out}(\Upsilon)) = \text{Val}(\text{Out}(\Upsilon'))$$

Both being maximal and order-invariant are typical properties of COORDINATION CASES. While I indeed restrict this investigation to COORDINATION CASES with these properties, not all COORDINATION CASES necessarily have these properties.<sup>109</sup>

---

<sup>109</sup>An example without these properties can be easily construed, at least if we allow a certain kind of case distinction. Consider

**Case 5.1 MUTUAL TERMINATION** John has kidnapped Adam and Lawrence. Lawrence wakes up in a small room with a timer showing 10 minutes and a big red button. Adam wakes up simultaneously in another room with the same setup. If both wait ten minutes without pressing their button, the doors open, and both stay alive. However, they also receive no additional reward. If Adam pushes his button first, he gets free and gets one million dollars. However, this floods Lawrence's room with a poisonous gas that kills him cruelly within seconds. In contrast, if Lawrence pushes first, he kills Adam through toxic gas and gets free, but Adam would not get any reward. If Adam pushes first and Lawrence, in his death throes, also pushes his button, not only will Adam also die, but John will next kidnap Adam's wife for one of his cruel experiments; but if Lawrence pushes first and Adam, in his death throes, also presses his button, this has no additional effect (beyond Adam's death). Similarly, in the improbable case of actual synchronous pushing, both die.

Here is a normal form representing the case:

		Lawrence	
		not-push	push
		not-push	push
Adam	not-push	both live on, no rewards (best)	only Lawrence lives on, but no reward (second-worst?)
	push	only Adam lives on and is rich (second-best)	both dead, but if Adam pushed first, his wife is kidnapped next (worst)

It is easy to see that MUTUAL TERMINATION is not maximal (since »combinations« of one agent's action are proper since not-pushing is by description meant to be a deliberate action, but one that is not necessary to »resolve« the situation). Nor is it order-invariant (in the case where both push their buttons, it becomes important – for Adam's wife, at least – whether Adam pushed first).

At this point, it is useful to introduce the notions of an *equivalence relation* and of *equivalence classes* for combinations of actions relative to a decision situation and a valuation function, which will prove useful later. Let  $D$  be a collective decision situation with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$ , and let  $\text{Val}$  be a valuation function. Then the relation  $\sim \subseteq \Psi \times \Psi$  is defined for combinations  $\Upsilon, \Upsilon' \in \Psi$  as

$$\Upsilon \sim \Upsilon' \text{ if and only if } \text{Val}(\text{Out}_C(\Upsilon)) = \text{Val}(\text{Out}_C(\Upsilon'))$$

is obviously an equivalence relation.<sup>110</sup> Further, we define the equivalence class of a combination as

$$[\Upsilon] := \{ \Upsilon' \in \Psi \mid \Upsilon \sim \Upsilon' \}$$

These concepts connect to the property of order-invariance insofar that, for all order-invariant decision situations, it holds that

$$\forall \Upsilon, \Upsilon' \in \Psi : ((\forall \phi \in \Upsilon : \phi \in \Upsilon') \wedge (\forall \phi \in \Upsilon' : \phi \in \Upsilon)) \rightarrow \Upsilon' \in [\Upsilon]$$

In other words, if two proper combinations contain the same actions, then they lead to equally good outcomes.

Based on this notion, we can introduce a more fine-grained conception of equivalence classes that also take into consideration the combination-constituting actions and that I later use for reducing the state space of *generalized extensive forms*, i.e., a unifying representation for COORDINATION CASES and SEQUENTIAL CASES. Let  $D$  be a collective decision situation with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$ , and let  $\text{Val}$  be a valuation function. Then

---

<sup>110</sup>This is true simply because, by assumption,  $=$  is an equivalence relation over the value space.

$\sim^* \subseteq \Psi \times \Psi$  as being defined, for two combinations  $\Upsilon, \Upsilon' \in \Psi$ , as

$$\Upsilon \sim^* \Upsilon'$$

if and only if

$$\text{Val}(\text{Out}_C(\Upsilon)) = \text{Val}(\text{Out}_C(\Upsilon')) \wedge (\forall \phi \in \Upsilon : \phi \in \Upsilon') \wedge (\forall \phi \in \Upsilon' : \phi \in \Upsilon)$$

is, again, an equivalence relation.<sup>111</sup> Further, we define the equivalence class of a combination as

$$[\Upsilon]^* := \{ \Upsilon' \in \Psi \mid \Upsilon \sim^* \Upsilon' \}$$

By construction,  $[\cdot]^*$  split the formerly  $[\cdot]$  equivalence classes further into parts by only letting combinations be in a common class if they have morally equivalent outcomes *and* involve the same individual actions.

Next, we define the set of equivalence classes for a given collective decision situation  $D$  with domain  $\Psi$  (relative to some valuation function) as

$$[\Psi]^* := \{ [\Upsilon]^* \mid \Upsilon \in \Psi \}$$

and accordingly we define *the set of the sets of representatives*

$$\mathcal{R} := \{ R \subseteq \Psi \mid \forall [\Upsilon]^* \in [\Psi]^* : \exists! \Upsilon \in R : \Upsilon \in [\Upsilon]^* \}$$

In other words, every  $R \in \mathcal{R}$  is a set that, for every  $[\cdot]^*$  equivalence class of combinations relative to  $D$  (and an arbitrary but fixed valuation function), contains *exactly one* element that belongs to that equivalence class. We can imagine constructing such a *set of representative*  $R \in \mathcal{R}$  by going through all equivalence classes and picking exactly one arbitrary element from it.

Now, we come to the final crucial property for this investigation.

---

<sup>111</sup>The reflexivity, symmetry, and transitivity of  $\sim^*$  should be obvious given that  $=$  is an equivalence relation over the value space and given that the universal quantifier together with the »is element of« relation cannot break the inherited properties. I leave the proof for the reader.

### 5.2.1.3 Symmetry

There is a third property that plays a crucial role in this part of my project: *symmetry*. Symmetry is a property of COORDINATION CASES.<sup>[112]</sup> Intuitively speaking, a COORDINATION CASE is symmetric if and only if it does not matter which agent performs which action; as long as they perform *corresponding* actions, the outcomes are identical with respect to their morally relevant quality. While the correspondence of the actions makes it somewhat challenging to capture this property formally, it is often intuitively obvious how the idea is meant. For instance, TWO FACTORIES is symmetrical: whether Ann pollutes and Ben produces cleanly or whether Ben pollutes and Ann produces cleanly, the outcomes are – qualitatively speaking – identical by construction. True, different workers are laid off, but the harm done to them is, by assumption, identical. In this case, the correspondence relation is implied by the type of the options.<sup>[113]</sup>

However, if we would – for what reason whatsoever – think of Ann’s option of pollution as corresponding to Ben’s option of producing cleanly and vice versa, we might think that the situation was *not* symmetrical. To illustrate this, let us assume that we number the agents’ options and give corresponding options the same number. If we then assume that in the normal form, we list the options of the agents according to their number in ascending order – from top to bottom and from left to right –, the intuitive correspondence mapping

<sup>[112]</sup>There is an interesting generalization of symmetry to SEQUENTIAL CASES, but I leave this point for another occasion as it does not fit into the scope of my thesis.

<sup>[113]</sup>In many well-known examples, the situation is similar. This is explained by the fact that it makes the examples easier to describe. However, it is by no means a necessary condition for TROUBLEMAKER (I will provide an example of this later in this section). But it is certainly not far-fetched that the fact that it behaves just like that is the best explanation that also Kagan was referring several times to actions »of a given kind« (cf. Section 4.4.1).

of Two FACTORIES would give our well-known normal form

		Ben	
		pollute	produce cleanly
		pollute	second-worst
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

while the unintuitive mapping would give us this one

		Ben	
		produce cleanly	pollute
		pollute	worst
Ann	pollute	worst	second-worst
	produce cleanly	best	worst

Obviously, the first normal form is symmetrical in a *geometric* sense: the normal form could be ›mirrored‹ along the main diagonal, i.e., along the imaginary line from the upper left to the top right, without changing the normal form in a morally relevant way (at least from a consequentialist point of view). This is not true for the second normal form (because both polluting has worse consequence than both producing cleanly).

While in Two FACTORIES the ›correct‹ representation seems obvious – induced by the type of the agents' options, so to speak – this is not necessarily so, as the following example shows. Consider

**Case 5.2 (LUCKY LISA)** *Johnny is a doctor, and he is really bad at his job. When Lisa asked him for some painkillers against her migraine, he accidentally gave her drug X, which, under normal circumstances, would cause Lisa tremendous pain, pain much worse than her migraine. However, drug X also makes*

*her immune against truth serum Y for several hours. At the same time, getting injected with that serum would also neutralize the painful effects of drug X.*

*After leaving Johnny's practice, Lisa takes her dose of drug X on her way home, where Mark is already waiting for her, hiding behind her front door. Mark wants to squeeze some secret information out of Lisa with the help of truth serum Y. This information would allow Mark to blackmail Lisa in the future, inflicting as much disutility on her as drug X's default pain effect does. After overpowering Lisa and tying her up to a kitchen chair, he injects truth serum Y into her. Thanks to drug X, Lisa can keep her secret and lie to Mark so that he ultimately fails with his evil plans to blackmail Lisa and, at the same time and unintendedly, he neutralizes drug X's painful effect. Still, Lisa suffered from a migraine for the rest of the day.*

One way to represent LUCKY LISA in a normal form is this

		Johnny	
		give drug X	give pain killers
		inject serum Y	second-worst
Mark	inject serum Y	second-worst	worst
	don't inject serum Y	worst	best

Obviously, our two protagonists, Johnny and Mark, are part of a TROUBLEMAKER that is structurally equivalent to Two FACTORIES. This normal form is symmetric for the same reasons. However, consider this alternative representation of LUCKY LISA:

		Johnny	
		give pain killers	give drug X
Mark	inject serum Y	worst	second-worst
	don't inject serum Y	best	worst

Geometrically speaking, this form is *not* symmetrical. However, unlike with TWO FACTORIES, I cannot really see why one of the two normal forms should be more representative than the other. I think they are on par.

The right reaction to this is not to introduce a convoluted theory with respect to correct or adequate correspondence relations. This investigation won't take a stand on the question of what makes correspondence mappings adequate or even whether this very idea actually makes sense in a general setting or proves theoretically fruitful. Instead, let us call all decision situations symmetric for which there is at least one symmetric normal form. Here is a more precise<sup>114</sup> definition:

**Property 5.3 (Symmetry)** *Let D be a maximal, collective decision situation with domain Ψ and two agents A<sub>1</sub> and A<sub>2</sub> with corresponding option spaces*

$$\Phi_{A_i} = \{\phi_{A_i}^1, \phi_{A_i}^2\}$$

*for i ∈ {1, 2}. D is symmetric if and only if there is at least one mapping between the Agents' option spaces such that, for every proper combination of actions, the outcomes are valuative equivalent to the combination that results from the original combination by applying that mapping, i.e., there is a permutation τ : I<sub>2</sub> → I<sub>2</sub> (i.e., τ ∈ S<sub>2</sub>) such that, for i, j, k, l ∈ {1, 2} and k ≠ l it holds that*

$$\forall \langle \phi_{A_k}^i, \phi_{A_l}^j \rangle \in \Psi : \text{Val}(\langle \phi_{A_k}^i, \phi_{A_l}^j \rangle) = \text{Val}(\langle \phi_{A_k}^{\tau(j)}, \phi_{A_l}^{\tau(i)} \rangle).$$

<sup>114</sup>For simplicity's sake, I restrict it to maximal COORDINATION CASES involving two agents with the same number of options.

In other words, we can find an order of the actions of one agent such that they line up to a symmetric normal form.

It is easy to mix up order-invariance and symmetry. To highlight the difference between the two, consider the following case that is order-invariant but not symmetric:

**Case 5.3 (HENRY’s HARDSHIP )** *In the wake of the recent recession, Henry has lost his apartment. Henry’s friend Molly has a vacant but unfurnished room in her apartment and could give it to Henry for free. Rico also had a hard time and had to move into a tiny apartment that is now filled with all the excess furniture he wants to get rid of. When Rico gives Henry the furniture but Molly doesn’t offer him the room, it doesn’t help Henry, in fact, it makes things a little worse: either Henry accepts the furniture nevertheless, so that he is homeless and has to take care of his cumbersome new possessions. Or he rejects Rico’s offer, which would make him angry. Regardless of whether Rico gives him the furniture, Henry will be better off if Molly offers him the spare room. Of course, Henry would be best off if he got the room and the furniture.*

This case can be represented in terms of four normal forms (cf. Table 5.1). Obviously, none of these normal forms is symmetric and, hence, HENRY’s HARDSHIP is not symmetrical. Yet, the case, like all COORDINATION CASES, is order-invariant.

TWO FACTORIES, however, is both order-invariant and symmetric. We can easily see how these properties interlink to an overall interesting structure. That TWO FACTORIES is order-invariant means, among other similar equalities, that

$$\langle \text{pollute}_{\text{Ann}}, \text{produce cleanly}_{\text{Ben}} \rangle$$

		Rico		Rico	
		give furniture away	don't give furniture away	don't give furniture away	give furniture away
Molly	offer room	best	second-best	offer room	second-best
	don't offer	worst	second-worst	don't offer	second-worst
		Rico	don't give furniture away	Rico	give furniture away
Molly	give furniture away	give furniture away	don't give furniture away	don't give furniture away	give furniture away
	don't offer	worst	second-worst	don't offer	second-worst
Molly	offer room	best	second-best	offer room	second-best

**Table 5.1:** The four normal forms of HENRY'S HARDSHIP . Note that we get four and not two (as one might expect in light of Property 5.3 as  $|S_2| = 2$ ) since we can decide on two different ways to enumerate the first agent's options. Thus, we get two times two normal forms. Considered crosswise (on the metalevel), the corresponding normal forms are strictly equivalent in terms of symmetry since the same options end on the main diagonal. Thus, the formulation of Property 5.3 is >sufficiently general<, i.e., it covers all cases.

has valutative identical outcomes as

$$\langle \text{produce cleanly}_{\text{Ben}}, \text{pollute}_{\text{Ann}} \rangle.$$

However, that Two FACTORIES is symmetric means that

$$\langle \text{pollute}_{\text{Ann}}, \text{produce cleanly}_{\text{Ben}} \rangle$$

has valutative identical outcomes as

$$\langle \text{pollute}_{\text{Ben}}, \text{produce cleanly}_{\text{Ann}} \rangle.$$

As we have seen in the first part, the typically discussed TROUBLEMAKERS that are COORDINATION CASES are maximal, order-invariant, and symmetric, i.e., cases satisfying the TRIAD. Accordingly, as announced, I restrict my investigation primarily to COORDINATION CASES with these properties. Interestingly, for every >triadic< COORDINATION CASE with domain  $\Psi$  which is symmetric to some correspondence mapping  $\tau$  it holds

that for any combination  $\langle \phi_{A_1}^i, \phi_{A_2}^j \rangle \in \Psi$  :

$$\begin{array}{ccc} \text{Val}(\langle \phi_{A_1}^i, \phi_{A_2}^j \rangle) & = & \text{Val}(\langle \phi_{A_1}^{\tau(i)}, \phi_{A_2}^{\tau(j)} \rangle) \\ \parallel & & \parallel \\ \text{Val}(\langle \phi_{A_2}^j, \phi_{A_1}^i \rangle) & = & \text{Val}(\langle \phi_{A_2}^{\tau(j)}, \phi_{A_1}^{\tau(i)} \rangle) \end{array}$$

The horizontal equivalences are guaranteed by *symmetry*, while the vertical equivalences are warranted by *order-invariance*. That all these combinations are proper is given by *maximality*.

### 5.2.2 1-Variants and Independence

Finally, note that every maximal COORDINATION CASE is, by definition, act-independent in the sense of INDEPENDENCY OF ACTION as introduced earlier. Recall:

**Property 3.9 (INDEPENDENCY OF ACTION)** *Let  $D$  be a collective decision situation. A combination of actions that is proper within  $D$  is act-independent if and only if any unilateral deviation of the combination is also proper in  $D$ .  $D$  is act-independent if and only if all combinations of actions are act-independent.*

After all, for every combination  $\Upsilon$  in such  $\Psi$ , we can substitute any action of any agent with another option of that agent without leaving the domain, i.e., we get another combination that is also proper. The following notion captures this operation of unilateral modification of combinations:

#### Definition 5.2 (1-Variation of a Combination)

*Let  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle$  be a collective decision situation. Let  $\Upsilon \in \Psi$  be an arbitrary proper combination of actions.  $\Upsilon'$  is a 1-variation of  $\Upsilon$  if and only if  $\Upsilon'$  differs from  $\Upsilon$  with respect to exactly one action of one agent.*

We can now restate the above thought: Given that a case is maximal, we can be sure that every 1-variant of a proper combination within that case is also proper. Without the assumption of maximality and, thus, INDEPENDENCY OF ACTION, the relevant notion – the idea of *minimal* variants of combinations – becomes much harder to capture but allows for much more sophisticated acceptable dependencies of actions in TROUBLEMAKERS, i.e., for a higher level of generality<sup>115</sup>

Based on the notion of 1-variation, we can straightforwardly define the *set of 1-Variants*  $\Psi_{\Gamma}^{\gamma,1}$  for an arbitrary combination of  $n$  actions  $\gamma \in \Psi_{\Gamma}$  relative to a domain  $\Psi_{\Gamma}$  over a set of option spaces  $\Gamma$ . We define

$$\Psi_{\Gamma}^{\gamma,1} := \{ \gamma' \in \Psi_{\Gamma} \mid \exists! \phi_A \in \gamma : \phi_A \notin \gamma' \wedge \exists! \phi'_A \in \Phi_A : \phi'_A \in \gamma' \}$$

This notion will make capturing the formal structure of TROUBLEMAKERS later quite easy. Again, when obvious, we will drop the sometimes unnecessary index  $\Gamma$  and will simply write  $\Psi^{\gamma,1}$ .

### 5.3 Revisiting SEQUENTIAL CASES

While I restricted the above considerations to COORDINATION CASES until now, it is finally time to turn to SEQUENTIAL CASES again. Since SEQUENTIAL CASES are defined as collective decision situations in which the order is specified, it cannot surprise that SEQUENTIAL CASES are normally *not* order-invariant. Consider JOB MARKET. The combination of Paul accepting the job first and George declining it second makes no sense, given the case description. The case is also not maximal: George's action of accepting the job alone suffices to specify an outcome.

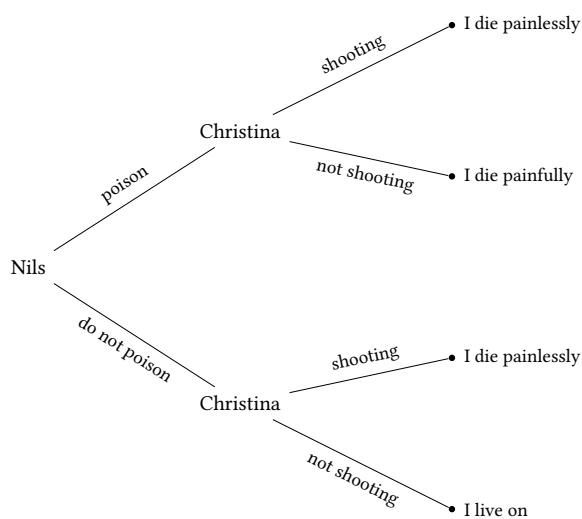
---

<sup>115</sup>This was already indicated in Footnote 85. I leave these details for another occasion.

Obviously, a **SEQUENTIAL CASE** can be maximal. Consider the following case<sup>116</sup> inspired by Derek Parfit (1984, p.70):

**Case 5.4 (DEADLY EVENING)** *Nils could trick me into drinking some poison of a kind that causes a painful death within a few minutes. Before this poison takes effect, Christina can kill me painlessly by shooting me.*

The extensive form of DEADLY EVENING is given by



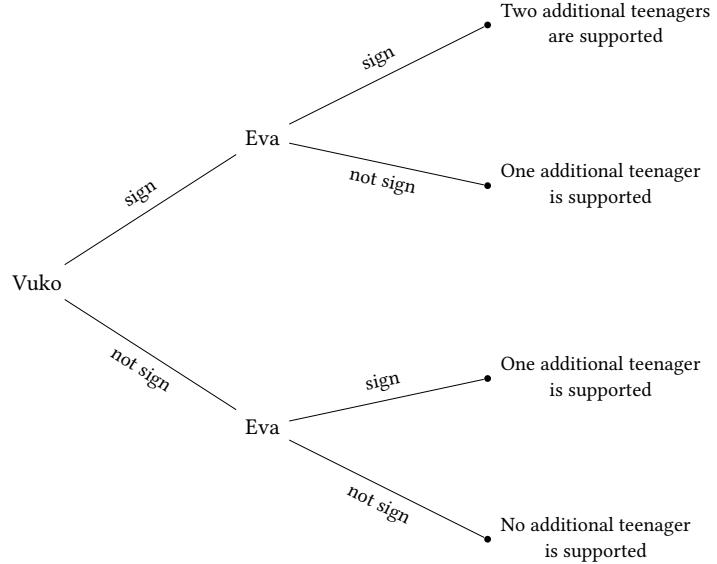
DEADLY EVENING, no doubt, is an exemplar of a *maximal* **SEQUENTIAL CASE** (and even a **THRESHOLD CASES** with threshold size of one).

Similarly, **SEQUENTIAL CASE** can even be order-invariant, at least in a slightly different sense than introduced above. Consider this case for example:

**Case 5.5 (CHARITY)** *Tina collects signatures for a charity project and goes from house to house. For each signature, the city sponsors one socially disadvantaged teenager with a scholarship. Vuko and Eva live in the last two houses on Tina's tour.*

<sup>116</sup>We can also refer to the modified version of JOB MARKET on page 79 where I have already indicated the following case.

CHARITY can be represented by the following extensive form



One could argue that CHARITY is order-invariant insofar that the outcomes *would* be equally good even if Tina first rang at Eva's door and then at Vuko', no matter what both decide to do. However, as the order of decisions is *fixed* to first Vuko, then Eva, these alternative order combinations are improper. Further, SEQUENTIAL CASES that are, in that *counterfactual sense*, order-invariant are ›boring‹ as they can be (losslessly) separated into two qualitatively identical, independent individual decision situations – which will be demonstrated in a moment. So, while order-invariance is not a truly interesting property with respect to SEQUENTIAL CASES, at least in the context of the CHALLENGE, the question of what makes a collective decision situation *separable* in that sense, is crucial for this project.

## 5.4 Separability and Conditionalization

There are at least two ways to ›derive‹ individual decision situations from collective ones. The first one I will call *decomposition*, and the second I call

*reduction* (through conditionalization). Decomposition does not seem applicable to all collective situations, but some seem decomposable. This speaks in favor of the GENUINE KIND VIEW and against COMPOSITIONALISM.

Recall these two conflicting views:

**Claim 5.1 (COMPOSITIONALISM)** *All collective decision situations can be reduced to individual decision situations (plus some structure).*

**Claim 5.2 (GENUINE KIND VIEW)** *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

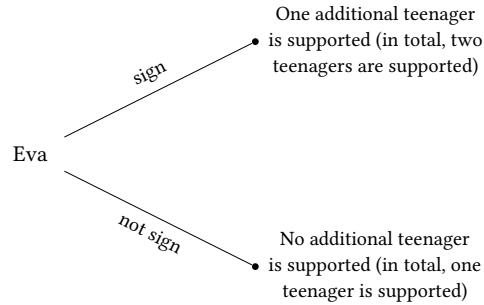
I shall call collective decision situations that are decomposable *separable*. I claimed above that CHARITY is an example of a separable collective decision situation. Indeed, it can be decomposed in individual decision situations without losing morally relevant information, at least from a consequentialist perspective: we can assess Eva's decision situation independently from Vuko's action. DEADLY EVENING, however, is an example for a *non-separable* situation: we cannot assess Christina's decision independently from Nils' action. These claims certainly need some support, and I will use the technique of *conditionalization* as introduced in Chapter 3. Recall:

**Principle 3.10 (CONDITIONALIZATION)** *Let  $C$  be a context, and let  $F$  be some state of affairs. If it is true, relative to  $\llbracket C \oplus F \rrbracket$ , that  $p$ , then it is true, relative to  $C$ , that [iff, then  $p$ ].*

Recall that » $\llbracket C \oplus F \rrbracket$ « is short for » $C$  together with the assumption that  $F$  obtains«, i.e., it expressed that, assuming  $C$  to be the actual context of some decision situation  $D$ , the circumstances of decision situation  $D$  are *extended* by  $F$ . Let us use this idea to CHARITY first. Assume:

**Fact 5.1** *Vuko signs the petition.*

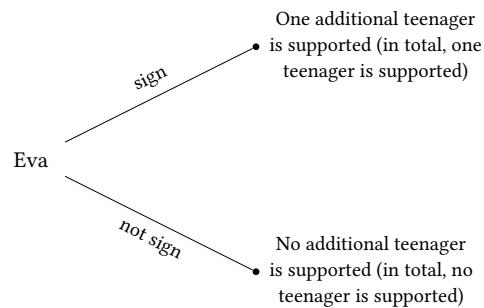
If we combine the context implied by the description of CHARITY with Fact 5.1, then arguably, there remains an individual decision situation for Eva, namely:



This is what I call a *structural decomposition of a collective decision situation by conditionalizing over some action* – in this specific case, a decomposition of CHARITY by conditionalizing over the action as referred to by Fact 5.1. Alternatively, we can conditionalize over

**Fact 5.2** *Vuko does not sign the petition.*

Then we get:



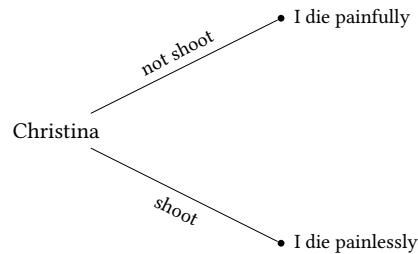
Let's assume that each supported teenager benefits equally from the scholarship. Then, the gain in value that Eva can contribute by signing (or that Eva can block by refusing to sign, respectively) is the same in both cases. This means that what Eva can contribute, morally speaking, is *independent*

from what Vuko does. Accordingly, we can assess Eva's situation without considering Vuko's action. This is, we can decompose CHARITY into two valuatively identical individual decision situations for Eva, resulting from conditionalizing over Vuko's two actions. Note that the same is true *vice versa*: We can assess Vuko's decision without considering Eva's action. To that extent, there is no particular reason to consider CHARITY in its entirety, as we could also consider the decision situations of both agents separately without losing anything.

Compare this with DEADLY EVENING. First, consider:

**Fact 5.3** *Nils poisons me.*

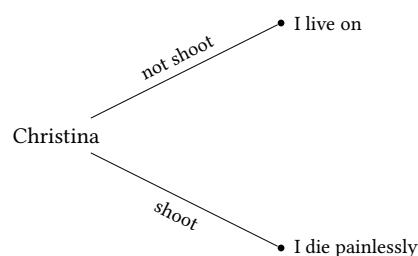
This reduces DEADLY EVENING to:



In this case, arguably, it is better for me if Christina shoots me. But now consider:

**Fact 5.4** *Nils does not poison me.*

This leaves Christina with:



No doubt that I am far better off in this case if Christina does *not* shoot me. Therefore, in moral terms, Christina's decision situation depends crucially on what Nils does. Accordingly, we cannot assess her situation without considering Nils' actions. This difference is what makes situations like CHARITY *reducible* but DEADLY EVENING *not*. We can extract individual decision situations in the described sense, which ultimately allows us to decompose the collective decision situation that we can assess in isolation by applying MOCOR. For irreducible situations like DEADLY EVENING, this cannot be done. All we get are conditional assessments by applying CONDITIONALIZATION.

This distinction is related to the so-called SURE-THING PRINCIPLE<sup>[117]</sup>. Translated to the moral, consequentialist domain, the SURE-THING PRINCIPLE suggests, roughly, that when evaluating potential outcomes relative to certain indeterminacies, these indeterminacies can be disregarded if the outcomes are occurring regardless of how those indeterminacies unfold. We can capture the SURE-THING PRINCIPLE in similar terms as we did with CONDITIONALIZATION:

**Principle 5.2 (SURE-THING PRINCIPLE)** *Let  $C$  be a context, and let  $F$  be some fact that will either obtain or not (which we refer to as  $\neg F$ ). If it is true, relative to  $\llbracket C \oplus F \rrbracket$ , that  $p$ , and it is true, relative to  $\llbracket C \oplus \neg F \rrbracket$ , that  $p$ , then it is true, relative to  $C$ , that  $p$ .*

---

<sup>117</sup>The *locus classicus* for the SURE-THING PRINCIPLE certainly is L. J. Savage (Savage 1954). Richard Jeffrey (Jeffrey 1982) and, much more recently, Judea Pearl (Pearl et al. 2000) highlighted the importance of certain independence assumptions between cause/action and effect given the background model. But thanks to the fact that we have agreed to presuppose such kind of independence in the cases we consider, this does not need to bother us. The SURE-THING PRINCIPLE has been contested by Colin Blyth (Blyth 1972), but his attack based on Simpson's Paradox (see Simpson 1951) violates the independence assumption as well (see Pearl 2016).

We can now state the distinction between separable and non-separable collective decision situations like this:

**Definition 5.3 (SEPARABILITY)** *Let  $D$  be a collective decision situation of  $n$  agents.  $D$  is reducible if and only if CONDITIONALIZATION allows us to decompose  $D$  into a class of evaluatively identical individual decision situations for each agent.*

Note that the SURE-THING PRINCIPLE gives us good reasons to adopt SEPARABILITY. Because, given that a collective decision situation is separable according to SEPARABILITY, the SURE-THING PRINCIPLE allows us to infer non-conditional assessments for all  $n$  agents within that situation. However, the other direction is not true: Not every situation allowing unconditional assessments is separable in that sense. After all, what needs to be constant over the results of conditionalizations to allow for unconditional assessments is only the *ranking* of outcomes, but not the exact evaluative profile (as demanded by SEPARABILITY). Recall HENRY'S HARDSHIP and the corresponding normal form:

		Rico	
		give furniture away	don't give it away
		best	second-best
Molly	offer room	best	second-best
	don't offer	worst	second-worst

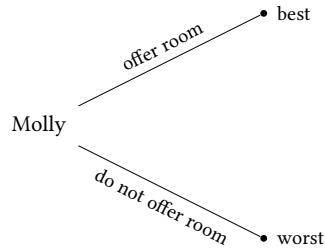
Trivially, CONDITIONALIZATION and the SURE-THING PRINCIPLE can be applied also to COORDINATION CASES. For instance, relative to HENRY'S HARDSHIP it is true that

- (25) It is better if Molly offers the free room to Henry.

because no matter what Rico does or will do, it is better for Henry not to be homeless. But the outcomes still differ morally. Conditionalizing over

**Fact 5.5** *Rico gives his spare furniture to Henry.*

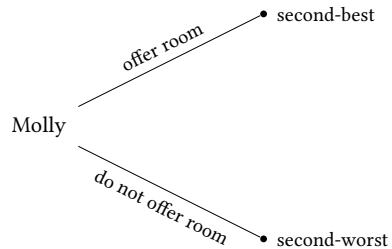
leaves Molly with the individual decision situation



while conditionalizing over

**Fact 5.6** *Rico does not give his spare furniture to Henry.*

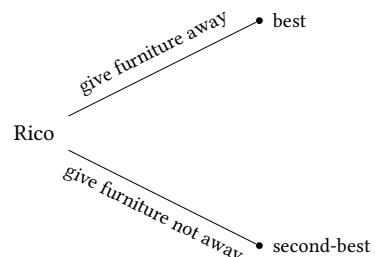
leaves Molly with



However, for Rico, we only get conditional assessments by conditionalizing over Molly's potential actions. Because conditionalizing over

**Fact 5.7** *Molly offers her room to Henry.*

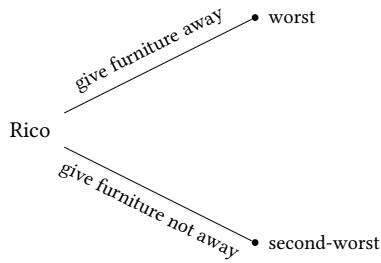
leaves Rico with the individual decision situation



while conditionalizing over

**Fact 5.8** *Molly does not offer her room to Henry.*

leaves Rico with

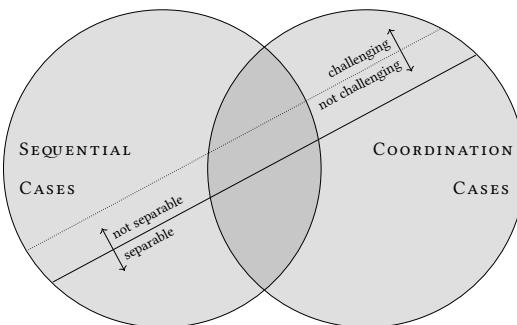


While in light of Fact 5.7, it would be right for Rico to give Henry his spare furniture, it is wrong for him to do so in light of Fact 5.8. For Rico, we can only get conditional assessments by applying CONDITIONALIZATION.

The example shows three things.

First, we can apply conditionalizing to COORDINATION CASES, not just to SEQUENTIAL CASES. Second, the possibility to arrive at unconditional assessments in a collective action situation with SURE-THING PRINCIPLE does *not* imply that the situation is separable. Sometimes, it is enough that

the moral ranking of the outcomes is good enough, as is the case for HENRY's HARDSHIP with respect to Molly. Such cases are to some extent *uninteresting* for this project, although they are *not* separable in the strict sense. Decomposing them implies losing some morally relevant information, even



**Figure 5.3:** The relation between the different sorts of collective decision situations. There are challenging (not-separable) and non-challenging (separable) cases of both kinds. Furthermore, all challenging cases are non-separable, but there are non-challenging cases that are non-separable, e.g., HENRY's HARDSHIP .

if not enough to affect moral assessments in terms of MOAC. Third, a collective action situation may be uninteresting with respect to one agent and interesting with respect to another precisely because it does not allow unconditional assessments for that agent. This is the case, for example, with respect to Rico in HENRY’s HARDSHIP .

Summing up, separable situations, thus, are only a specific class of what rightfully might be called *non-challenging* (collective decision) situations, i.e., of collective decision situations that pose no particular challenge to MOAC. With respect to the COMPOSITIONALISM versus GENUINE KIND VIEW question we started this section with, separable cases do not matter. Figure 5.3 gives an overview of the different kinds of cases introduced so far.

### 5.4.1 Formal Toolbox for Reductions

For the remainder of this part, having clear semantics for conditional assessments proves very useful. Thus, we ensure that reductions can be formally captured in the proposed collective decision-making framework. In order to do so, we first define the following operator:

$$\Upsilon \ominus \phi := \begin{cases} \langle \psi_1, \dots, \psi_{i-1}, \psi_{i+1}, \dots, \psi_n \rangle & \text{if } \Upsilon = \langle \psi_1, \dots, \psi_{i-1}, \phi, \psi_{i+1}, \dots, \psi_n \rangle \\ \Upsilon & \text{otherwise} \end{cases}$$

In other words, the operator  $\ominus$  removes an option  $\phi$  from a combination if the option is part of the combination. If not, it leaves the combination unaffected.

Equipped with  $\ominus$ , it is easy to neatly define the reduction of collective decision situations, which we have already casually practiced many times in the last section. Let  $D$  be some collective decision situation  $D$  with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$ . We write  $D_{\downarrow \phi}$  for the reduction of  $D$  by conditional-

izing over some option  $\phi$  from some option space  $\Phi$  from  $\Gamma$  corresponding to the options of an agent  $A \in \mathcal{A}$ . Then  $D_{\downarrow\phi}$  is defined as

$$D_{\downarrow\phi} := \langle \mathcal{A}_{\downarrow\phi}, \Gamma_{\downarrow\phi}, \text{Out}_{[C \oplus \phi]} : \Psi_{\Gamma_{\downarrow\phi}} \rightarrow \mathcal{O} \rangle$$

with

$$\mathcal{A}_{\downarrow\phi} := \mathcal{A} \setminus \{A\}$$

$$\Gamma_{\downarrow\phi} := \Gamma \setminus \{\Phi\}$$

$$\Psi_{\Gamma_{\downarrow\phi}} := \{ \Upsilon \ominus \phi \mid \Upsilon \in \Psi_\Gamma \text{ with } \phi \in \Upsilon \}$$

$$\text{Out}_{[C \oplus \phi]}(\Upsilon_{\downarrow\phi}) := \text{Out}_C(\Upsilon) \text{ with } \Upsilon_{\downarrow\phi} := \Upsilon \ominus \phi$$

We can reiterate this operator for collective decision situations with more than two agents. This allows us to reduce *arbitrary* collective decision situations of  $n$  agents by conditionalizing over the actions of  $n - 1$  agents to an individual decision situation of the remaining agent. For this, we first define a *proper part* of a combination of actions  $\Upsilon$  to be a combination of actions  $\Upsilon'$  – in symbols:  $\Upsilon' \sqsubset \Upsilon$  – such that  $|\Upsilon'| < |\Upsilon|$  and  $\forall \phi \in \Upsilon' : \phi \in \Upsilon$ . For two combinations with  $\Upsilon' \sqsubset \Upsilon$  we define  $\Upsilon \ominus \Upsilon'$  to be the *missing part* of  $\Upsilon$  with respect to  $\Upsilon'$ , i.e., if  $\Upsilon \ominus \Upsilon' = \Upsilon''$ , then  $\Upsilon' \oplus \Upsilon'' = \Upsilon$ .

Now we can define  $D_{\downarrow\Upsilon'}$  for the reduction of  $D$  by conditionalizing over some proper part  $\Upsilon'$  of a proper combination  $\Upsilon \in \Psi$  of  $D$  as the repeated application of the above-defined reduction in the correct order:

$$D_{\downarrow\Upsilon'} := \left( \dots (D_{\downarrow\Upsilon'_1})_{\downarrow\Upsilon'_2} \dots \right)_{\downarrow\Upsilon'_{|\Upsilon'|}}$$

This implies that we can reduce any arbitrary collective decision situation (with a finite number of agents) to some *individual* decision situation, namely by reducing for a proper part  $\Upsilon'$  of a proper combination  $\Upsilon$  with  $|\Upsilon'| = |\Upsilon| - 1$ .

This allows me to extend my definition of the rightness property to finally >unlock< conditional assessments formally. That is, we want to define what it means, given a *collective* decision situation  $D$ , to say that

- (26) If the partial combination of actions  $\Upsilon'$  were realized, then it is right to  $\phi$  for  $A$ .

or, somewhat more naturally, given that  $\Upsilon' = \langle \phi_{A_{i_1}}, \dots, \phi_{A_{i_{n-1}}} \rangle$

- (27) If  $A_{i_1}$  performs  $\phi_{A_{i_1}}$  and ... and  $A_{i_{n-1}}$  performs  $\phi_{A_{i_{n-1}}}$ , then it is right to  $\phi$  for  $A$ .

So, let's  $T, D, C \vDash R(\phi | \Upsilon')$  express that proposition (26) and let  $D$  be a collective decision situation with actual context  $C$ . We define:

$$T, D, C \vDash R(\phi | \Upsilon') \quad \text{if and only if} \quad T, D_{\downarrow \Upsilon'}, [[C \oplus \Upsilon']] \vDash R\phi.$$

This allows us to express the **SURE-THING PRINCIPLE** (restricting to a case with two agents  $A_1$  and  $A_2$  with two options  $\phi_i$  and  $\neg\phi_i$  (for  $i \in \{1, 2\}$ ) for the sake of readability): if  $T, D, C \vDash R(\phi_1 | \phi_2)$  and  $T, D, C \vDash R(\phi_1 | \neg\phi_2)$ , then  $T, D, C \vDash R\phi_1$  (and analogously for  $\neg\phi_1$  and vice versa).

Applying all this to our running example, **TWO FACTORIES**, we thus can decide that the intuitive assessments of MOCOR, which played a central role in the reconstruction in Section 3.5.2.3 really have a solid basis in carefully applied theory. For instance,

- (28) If Ben pollutes, it is right for Ann to pollute.

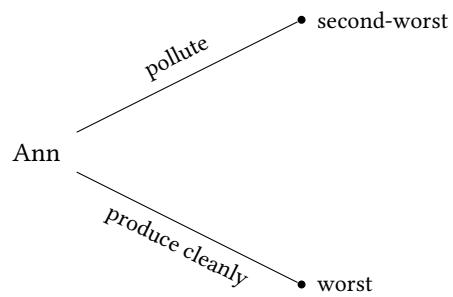
is true in **TWO FACTORIES** because

- (29) It is right for Ann to pollute.

is true in Two FACTORIES combined with the fact that

**Fact 5.9** *Ben pollutes.*

which results in the individual decision situation



After all, in this situation, the best that Ann can do is to pollute.

We now have a solid foundation for the semantics of such conditional assessments that arguably are ascriptions of certain *conditional deontic statuses*.<sup>118</sup> With this, I have filled the toolbox with the tools I need for this part. We can turn now to a deeper and better analysis of the CHALLENGE in all its facets.

---

<sup>118</sup>In our everyday language there are certainly different kinds of »Iffy Oughts«, as Fred Feldman called them (cf. chapter 4 in Feldman 1980). The ones introduced here are the kind relevant to the CHALLENGE, and I do not enter the field any further (but I claim that they bear an important structural resemblance to the conditional oughts Jeff Harty has explored in chapter 5 of Harty 2001).



# **Chapter 6**

## **The REAL CHALLENGE**

I begin the development of my own approach by dismissing the CHALLENGE<sub>int</sub>. For this, we revisit the ARGUMENT and give reasons for why it is invalid.

As we will see, this result helps proponents of MOAC much less than one might think initially because a new challenge arises from this insight, which gives this chapter its name: the REAL CHALLENGE. This challenge is more fundamental and is, in a sense, the *cause* of the CHALLENGE, which is thus more of a *symptom*.

The general insight is that because of the apparent inseparability of TROUBLEMAKERS, consequentialists decided for the argumentative move discussed in quite some detail in Section 3.5.2.3 to add facts about what other agents do, will do, or have done to the actual context of these collective decision situations. I shall call this the CONSEQUENTIALIST STANDARD MOVE (or, shorter, just the CSM).

The essential point of this chapter within the context of this project, then, is this: to avoid a fundamental challenge, consequentialism took a wrong turn – the CSM – and, as a result, ended up with the CHALLENGE in the first place.

However, without the CSM, camp MOAC must face a myriad of systematic deontic gaps, i.e., agree to an unbearable degree of deontic *incompleteness*. This is the **REAL CHALLENGE**, and it is why proving the **ARGUMENT** invalid in this sense is indeed a Pyrrhic victory for my consequentialist project. The task that remains, then, is to solve the **REAL CHALLENGE** without running into the **CHALLENGE** >again<.

Before turning to this task, however, I would like to briefly present one further observation as to why something seems wrong with the current treatment of collective decision situations, i.e., with the CSM. This should serve as rather independent motivation to fundamentally question the CSM that is independent of the **CHALLENGE** and, more specifically, from MH.

## 6.1 PRINCIPLE OF MORAL BALANCE

PMH is not the only principle threatened by collective contexts that proponents of MOAC arguably should care about. Here is another candidate that I have not been able to find in the literature. The basic intuition is that if two combinations of actions lead to morally equivalent consequences, the moral assessments of the individual actions making up the two combinations >must not diverge too much<. Call this general idea **PRINCIPLE OF MORAL BALANCE** (or PMB in short). Without PMB being true in some sense, the connection between the quality of consequences and the moral status of actions, which is of paramount importance to consequentialists, is undermined.

PMB leaves open what it means for a divergence to be too large deliberately. I think we can consider the widest possible divergence, which would occur when all actions that form one combination are deemed right and all

actions that make up the other combination are deemed wrong. This would be the most severe imbalance possible. Here is my suggestion for a concise formulation:

**Principle 6.1 (MORAL BALANCE (MB))** *Let  $D$  be a collective decision situation, and let  $\Upsilon$  and  $\Upsilon'$  be two proper combinations in  $D$ . If  $\Upsilon$  and  $\Upsilon'$  lead to morally equivalent consequences, then it cannot be that all  $\Phi \in \Upsilon$  are (necessarily) right and all  $\Phi' \in \Upsilon'$  are (necessarily) wrong.*

We have already seen a case where MB is apparently violated. Recall Glover's two BEANS AND BANDITS cases (cf. page 91). The first one is about bandits who obviously have never heard of the CHALLENGE (or just don't care if what they are doing is wrong or not, or who are not from camp MOAC ...), recall:

**Case 3.4 (BEANS AND BANDITS (One to One))** *Suppose that in a village, there are 100 very hungry, almost starving tribesmen who prepare their lunch on 100 small fireplaces. 100 mildly hungry bandits are waiting outside the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one villager's bowl to satisfy their appetite.*

There was no particular challenge in condemning these bandits' actions with MOCOR. Now recall the case where the bandits behave rather strangely (perhaps they have been tutored by a wandering practical philosopher in the meantime):

**Case 3.5 (BEANS AND BANDITS (Many to Many))** *Suppose that in a village, there are 100 very hungry, almost starving tribesmen who prepare their lunch on 100 small fireplaces. 100 mildly hungry bandits are waiting outside*

*the village for the right moment to steal the villagers' food. While the villagers are briefly distracted and turn their backs on their fireplaces, the bandits sneak into the village unnoticed. Each thief steals one bean from each villager's bowl to satisfy their appetite.*

With respect to PMB the following matters: Both combinations of actions lead to morally equivalent results, namely 100 hungry villagers without lunch and 100 well-satiated bandits. But in the one-to-one case, each of the bandits has done something wrong according to MOCOR (namely, harmed the corresponding villager by robbing them of their lunch), while in the many-to-many case, it seems that MOCOR finds that no one has done anything wrong. This violates MB. Note that this challenge is independent of PMH as it is not grounded in the fact that morally suboptimal results are produced in any way. Instead, it lies in the *divergence* of the respective moral assessments.

BEANS AND BANDITS is a CUMULATIVE EFFECTS CASE, and I have excluded these cases from this project (Section 4.2). But we can build COORDINATION CASES with similar structure:

**Case 6.1 (SCHOLARS' BIRTHDAY STANDOFF)** *Markus and Caro, both prominent figures in their scientific field, share the same birthday. On this particular day, they find themselves at an important scientific conference on a big and societally relevant topic together. For years, they have been at the center of a scholarly divide, each leading a camp of dedicated followers. The researchers present at the conference are distinctly aligned with either Markus or Caro, and both factions are roughly equal in size.*

*There are three possible options that Markus and Caro can choose from: Each of them could rise above their differences and wish the other a happy birth-*

*day. Alternatively, each could decide to avoid the topic of their shared birthday entirely. Finally, each could openly display their animosity.*

*From the perspective of conference harmony, mutual birthday wishes would foster the most positive atmosphere and lead to scientific breakthroughs that would help solve important global-scale challenges. If one extends birthday wishes and the other remains silent, the ambiance would still be reasonably relaxed. Minor progress would be made. However, if both avoid the birthday topic or show hostility, this would put a strain on the mood, but only slightly, as such behavior is in line with the expectations of the participants. The conference would be a lost opportunity, but in the medium term, the dust could settle. A one-sided show of hostility would create the most tension, degrading the conference's mood the most. This would also dampen spirits for many years to come and put a lasting brake on progress in the interests of society.*

SCHOLARS' BIRTHDAY STANDOFF is best conceived of as a COORDINATION CASE represented by the following normal form:

		Markus		
		congratulate	say nothing	be rude
		congratulate	say nothing	be rude
Caro	congratulate	+++	+	--
	say nothing	+	-	--
	be rude	--	--	-

There are two combinations with morally equivalent outcomes. Whether Markus and Caro simply exclude the topic of their birthday and remain silent about it or whether they are both rude does not matter morally. However,

we can apply the CSM to argue that in the first case (when both remain silent), they act wrongly according to MOCOR, but in the second case (when both are rude), they act rightly. Because given that both say nothing, it holds that Markus (Caro) could have made a difference for the better by acting differently, namely if he (she) had congratulated his (her) opponent. However, this is not true if both are rude to each other because then Markus (Caro) could not only have *not* made a difference for the better but even only for the worse by acting differently. If he (she) had congratulated her (him) or simply remained silent, it would only have led to further escalation and spoiled the mood of the conference altogether. Once again, the CSM seems to lead to assessments unacceptable for camp MOAC. This time, it is a violation of PMB and not PMH.

I do not want to go deeper into PMB or MB here. Instead, they should motivate us to question precisely this kind of (anticipatory) retrospective reasoning, i.e., the CSM, independent from the CHALLENGE itself. With this doubt in mind, we return to the CHALLENGE – and more specifically, the CHALLENGE<sub>int</sub> and the ARGUMENT, which rely heavily on this reasoning.

## 6.2 Revisiting the ARGUMENT

Recall, once again,

**Argument:** The ARGUMENT (tentative)

$P_{\exists T}$ : There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

$P_{MOACoR}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{MH}$ : If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

$C_{\neg Adeq}$ : MOAC is not an adequate moral theory.

As announced, my goal is to cast doubt on the deductive validity of the ARGUMENT. An argument is (deductively) valid if and only if the truth of its premises (deductively) entails the truth of its conclusion. In other words, it cannot be that the premises of a valid argument are true, but its conclusion is

false. Thus, to decide whether the ARGUMENT is valid, we must determine whether the conclusion follows (deductively) from its premises. Because the validity of an argument is a structural, semantic property of an argument, we should take a careful look at its semantic structure. We are thus searching for an adequate formal representation, which typically requires careful interpretation for complex arguments. Thankfully, we have already worked to understand the ARGUMENT (Section 3.5) and developed quite some formal toolbox (Chapter 5). However, before I turn to the formal representation of the ARGUMENT, I will highlight some essential but usually overlooked details and sketch the general intuition behind the interpretation I later advocate.

### 6.3 The Intuition: Gaps Filled Badly

Before I establish the invalidity of the ARGUMENT based on the formal toolbox introduced so far, I will briefly convey the general insight on which my reconstruction is based. First, we can assume, for now, the existence of TROUBLEMAKERS. Suppose the agents in some TROUBLEMAKERS realize a troublesome combination. Accordingly, they produce a suboptimal outcome (COLLECTIVE SUBOPTIMALITY), and, given their actions, none of the agents involved could have made a difference for the better by unilaterally acting differently (INDIVIDUAL OPTIMALITY). The CHALLENGE is based on two observations: on the one hand, in light of (INDIVIDUAL OPTIMALITY), MOCOR seems to yield that all agents have acted rightly; on the other hand, MH requires at least one wrong action in light of (COLLECTIVE SUBOPTIMALITY).

Second, we have understood that the CHALLENGE<sub>int</sub> rests on the two principles MOCOR and MH. If both MOCOR and MH are true, the matter

seems clear: MOAC is inadequate. Accordingly, it might be tempting to come up with a rough-and-ready formalization<sup>119</sup> of the ARGUMENT that looks like this:

**Argument:** The ARGUMENT (formally, naïve)

$$P_{\exists T}: \quad \exists D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi : \\ \neg \text{GlobOpt}(\Upsilon) \wedge \text{IndiOpt}(\Upsilon)$$

$$P_{\text{MOCoR}}: \quad \forall D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ \text{IndiOpt}(\Upsilon) \rightarrow \text{All-R}(\Upsilon, \text{MOAC})$$

$$P_{\text{MH}}: \quad \forall T : \text{Adeq}(T) \rightarrow \forall D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ \neg \text{GlobOpt}(\Upsilon) \rightarrow \neg \text{All-R}(\Upsilon, T)$$


---

$$C_{\neg \text{Adeq}}: \quad \neg \text{Adeq}(\text{MOAC})$$

If this were an adequate formal representation of the ARGUMENT, validity would be beyond question.  $P_{\exists T}$  together with  $P_{\text{MOCoR}}$  entails the existence of collective decision situations where a suboptimal combination of actions consists only of right actions, i.e.,

$$\exists D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi : \\ \neg \text{GlobOpt}(\Upsilon) \wedge \text{All-R}(\Upsilon, \text{MOAC})$$

---

<sup>119</sup>For the predicates  $\text{GlobOpt}(\dots)$  expressing that a combination of actions does produce the best consequences (and, thus,  $\neg \text{GlobOpt}(\dots)$ ) corresponding to COLLECTIVE SUBOPTIMALITY,  $\text{IndiOpt}(\dots)$  corresponding to INDIVIDUAL OPTIMALITY,  $\text{All-R}(\dots, \dots)$  expressing that all actions in a combination to be right according to some theory, and  $\text{Adeq}(\dots)$  expressing that some theory is adequate.

By exemplifying  $P_{\text{MH}}$  with MOCoR, we can then infer:

$$\begin{aligned} \text{Adeq}(\text{MOAC}) \rightarrow \exists D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi : \\ \neg \text{All-R}(\Upsilon, \text{MOAC}) \wedge \text{All-R}(\Upsilon, \text{MOAC}) \end{aligned}$$

Since the right side of the conditional is a logical falsehood – there can be no collective decision situation with a combination in which *all* actions are right and *not all* actions are right at the same time –, we get the result that MOAC is inadequate.

However, at a second glance and against the background of what we learned about the structure of collective decision situations in Chapter 5, the above reconstruction is overly simplistic. Most importantly, MOCoR only arrives at the alleged assessments if, in assessing each action, it is already considered as a matter of fact what all other agents have done. The assessment is made in *retrospect*. This is what CSM was all about. Adding facts about what the other agents do, will do, or have done to the actual context in order to ›escape‹ the conditional.

MH, however, relies on the plausibility of **COLLECTIVELY MAXIMIZING**. Recall:

**Principle 5.1 (COLLECTIVELY MAXIMIZING)** *Let  $D$  be a collective decision situation with domain  $\Psi$  and with actual context  $C$ . If  $\Upsilon \in \Psi$  consists only of right actions, then there is no (and cannot be) alternative  $\Upsilon' \in \Psi$  with better consequences than  $\Upsilon$  relative to  $C$ .*

**COLLECTIVELY MAXIMIZING** arguably has a kind of guiding, anticipatory, prospective character. Accordingly, in line with all that we have learned in the first part of this project, MH is meant to express the idea that morality

has the function of implicitly *coordinating* our behavior in such a way as to produce the best outcome we can collectively produce. A few witnesses shall be called again. First, here is Fred Feldman (1980, p. 168):

Quite a few moral philosophers seem to believe that when all the members of a social group do what they morally ought to do, the group as a whole does benefit more than it would have from the performance of any worse alternative set of actions. I shall say that any such view is a version of the Principle of Moral Harmony (PMH).

Similarly, but much more recently, Douglas Portmore (2018, p. 12) reads MH like this:

The principle of moral harmony holds that a moral theory must be morally harmonious – that is, it must be such that the agents who satisfy the theory, whoever and however numerous they may be, are guaranteed to produce the morally best world that they together have the option of producing.

Or remember Stephen Toulmin's dictum (1953, p.137) according to which »we can provisionally define [the ›function‹ of ethics] as being 'to correlate our feelings and behaviour in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible'.« It stands to reason that Toulmin here rather means that our behavior should be aligned, coordinated, attuned, ..., than ›retroactively assessed‹.

Thus, MH is first of all about where we arrive when we (with or without the intention to do so) *follow morality*, i.e., what we bring about when everything we do *satisfies* the moral theory under consideration (to use Regan's, Parfit's, and Portmore's lingo), if we select from the *set of right actions* (Re-

gan's lingo again), if ›doing what one ought‹ (Feldman) *according to* a theory, etc., etc. *This is not a backward-looking perspective.*

Therefore, to assess whether there are indeed violations of the MH in TROUBLEMAKERS, we should indeed proceed in the same way that Regan *attempted* to apply in the context of his impossibility result: We should look at each set of right actions for each agent implies by MOAC. If we then find a combination of actions consisting only of actions that were chosen in this manner, then surely we can rightly say that if the agents acted accordingly, then they are doing what they ought to do (in the sense of Feldman's quotation), respectively, and then they are fulfilling MOAC (in the sense of Protmore's quotation and also Regan's vocabulary). Suppose this combination of actions would also lead to suboptimal outcomes. In that case, we could rightly say that we have found a violation of COLLECTIVELY MAXIMIZING – and thus, according to MH, MOAC would be inadequate.

Thus, for troublesome combinations to be *truly* troublesome, they must consist only of right actions for the agents in a TROUBLEMAKER – period. That is, without adding facts about the other agents' actions. In other words, for the ARGUMENT to hold, there must exist right actions for the agents, i.e., TROUBLEMAKERS must be resolvable.

But can these forward-looking assessments really be warranted by or inferred on the basis of the backward-looking reasoning in  $P_{MOCoR}$ ? My claim is that they cannot because to apply MOCoR, we need to identify individual decision situations for all the agents. This, as was shown in the last chapter, can be done by reducing the collective decision situation by adding the actions of all other agents to the context of the collective decision situation, sure. But as we have seen, the so-derived assessments based on

the technique of conditionalizing only warrant conditional assessments with respect to the original collective decision situation. This is, we need to identify right actions for all agents in a TROUBLEMAKER *without* applying CSM.

As a result, then, even if, in retrospect, given the context that all agents have already acted, every action *was* right, this does not at all entail that none of the agents' actions *must have been* right at the relevant decision-time. Thus, the assessments referred to  $P_{MOCoR}$  are derived relative to another context than the assessments ›required‹ by MH. In this respect, the apparent trivial inference above does not do justice to the ARGUMENT.

Before I turn to a better reconstruction of the ARGUMENT, let me emphasize that the actions of other agents are not already part of the actual context. Assuming that they are is to assume determinism. But assuming determinism violates

**Principle 6.2 (METHODOLOGICAL INDETERMINISM)** *The question of what is right to do for an agent in a decision situation is pointless if what the agent does is already predetermined. Even if determinism were right, in the context of morality, we should pretend that it is not.*

I will not do much to defend this principle because I believe it to be quite self-evident. However, one quick reason for METHODOLOGICAL INDETERMINISM should be cited: We usually presuppose determinism to be false in our moral discourse. More specifically, for every decision situation, we assume that the agent populating that situation can perform any of their options and especially consequentialist discourse heavily relies on comparing possible consequences of options with other possible consequences of other options. But if we were to assess the options of all agents within a collective decision

situation, we need to assume the one agent's actions as determined in order to assess the other agent's options, and vice versa. So, we actually need to assume all agents' actions. There is no choice left that we could assess. I claim that all this gives a sufficiently strong reason for camp MOAC to embrace METHODOLOGICAL INDETERMINISM.

Naturally, I am not the first one to notice this struggle. Here is a passage<sup>120</sup> from A.N. Prior (1956, pp. 91-92, see also John F. Horty, 2001, chap. 4):

Suppose that determinism is *not* true. Then there may indeed be a number of alternative actions which we could perform on a given occasion, but none of these actions can be said to have any »total consequences«, or to bring about a definite state of the world which is better than any other that might be brought about by other choices. For we may presume that other agents are free beside the one who is on the given occasion deciding what he ought to do, and the total future state of the world depends on how these others choose as well as on how the given person chooses ; and even if there were not other people to spoil one's

---

<sup>120</sup>The context of Prior's quote is also interesting concerning METHODOLOGICAL INDETERMINISM. Prior starts from a point Moore made in his *Principia Ethica* (Moore 1903) concerning determinism and its importance for consequentialist considerations:

My argument is dilemmatic. Either determinism is true or it is not. If determinism is true then there are not really (though there may seem to be) a number of alternative actions which we could perform on a given occasion; the one action that we can perform is the one that we do perform. Hence whatever we in fact do is the best possible action (the one with the best possible total consequences) because it is the *only* possible action ; so that whatever we in fact do is our duty, is Moore's sense of »duty«. Moore himself saw this horn of the dilemma (and indeed it is a commonplace that determinism presents problems of this sort); but it has another horn which so far as I know he did *not* see. [..., the passage below]

The conclusion seems clear. If determinism is true, then whatever we do is our duty in Moore's sense of »duty«, and if determinism is not true then nothing at all is our duty in this sense.

The solution proposed in this part of my dissertation can thus be understood as a resolution of Prior's dilemma. It allows champions of MOAC to adopt METHODOLOGICAL INDETERMINISM, i.e., to work under the assumption of the falsity of determinism, without running into the problem raised here by Prior, which is ultimately the REAL CHALLENGE.

calculations there would still be oneself, with one's own future choices, or some of them, undetermined like this present one (unless a man decides that it is too risky for him to have any further freewill, and on this very ground finds it to be his duty to do away with himself). And while I speak here of one's calculations being spoilt, the trouble of course goes deeper than that – it's not merely that one cannot calculate the totality of what will happen if one decides in a certain way ; the point is rather that there *is* no such totality.

The conclusion seems clear. If [...] determinism is not true then nothing at all is our duty in this sense.

Call this challenge the **REAL CHALLENGE**. It is not that MOAC would make *incorrect assessments* in **TROUBLEMAKERS**, which then violates the ideal of PMH, but rather it is that MOAC simply *does not make any assessments at all* in these cases. This, of course, is also not in accordance with the spirit of PMH, but does not warrant any violations, though. But worse, MOAC, instead of having **DEONTIC COMPLETENESS**, is a Swiss cheese, full of systematic deontic voids, gaps, and holes.

All this might be a bit abstract, so we return to our running example, **TWO FACTORIES**, before turning to another reconstruction attempt. Recall the normal form:

		Ben	
		pollute	produce cleanly
	pollute	second-worst	worst
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

We now have a sufficient understanding of the situation to be sure that we can indeed derive all of the following propositions from **TWO FACTORIES**:

- (30) If Ann pollutes, it is right for Ben to pollute.
- (31) If Ben pollutes, it is right for Ann to pollute.
- (32) Once Ann has polluted, it will have been right for Ben to have polluted.
- (33) Once Ben has polluted, it will have been right for Ann to have polluted.
- (34) If Ann produces cleanly, it is right for Ben to produce cleanly.
- (35) If Ben produces cleanly, it is right for Ann to produce cleanly.
- (36) Once Ann has produced cleanly, it will have been right for Ben to have produced cleanly.
- (37) Once Ben has produced cleanly, it will have been right for Ann to have produced cleanly.

Similarly, relative to Two FACTORIES together with the fact that both pollute, i.e., that they instantiate the troublesome combination within Two FACTORIES, all of the following propositions are true:

- (38) That Ann polluted was right.
- (39) That Ben polluted was right.
- (40) That Ann polluted was right because Ben polluted.

(41) That Ben polluted was right because Ann polluted.

However, the following propositions are *false* relative to TWO FACTORIES together with the fact that both pollute

(42) That Ann polluted was right, independently from what Ben did.

(43) That Ben polluted was right, independently from what Ann did.

All this comes down to say that, while MOAC can give us quite a lot of assessments, it does *not* yield any non-conditional assessment relative to TWO FACTORIES, not even in retrospect – at least not relative to TWO FACTORIES without the actually performed actions, I mind you. This result is a structural one, and it can be generalized with respect to (many options of many agents within) non-separable collective decision situations and, thus, with respect to TROUBLEMAKERS. The problem, then, is not the CHALLENGE as there can be no violation of MH in light of non-resolvability; the problem with TROUBLEMAKERS is that MOAC is not able to give us any non-conditional assessments in such cases in the first place. I shall call this the REAL CHALLENGE.

At this point, one might object that all this is quite ›hand-wavy‹ – and rightly so. I owe the reader a better, careful, and formal reconstruction of the ARGUMENT.

## 6.4 The Logical Structure of The ARGUMENT

In this section, I will develop a reasonable formalization of the ARGUMENT. I turn to each of the three premises, one by one. Only then do I put them together and show the invalidity of the ARGUMENT.

### 6.4.1 The Structure of $P_{\exists T}$ : Straightforward

$P_{\exists T}$ 's structure is easily captured based on the semi-formalism introduced earlier. For this, first recall the tentative definition from the introduction:

**Definition 1.1 (TROUBLEMAKERS (tentative))** *A collective decision situation is a TROUBLEMAKER if and only if there is a troublesome combination of options therein, i.e., the agents can act in ways such that*

**COLLECTIVE SUBOPTIMALITY** *together they would produce a morally suboptimal outcome and*

**INDIVIDUAL OPTIMALITY** *none of them could make a difference for the morally better by unilaterally acting differently.*

Acknowledging that being a TROUBLEMAKER depends on axiological questions allows us to be slightly more precise or more general:

**Definition 6.1 (TROUBLEMAKER (relative to some axiology  $V$ ))** *A collective decision situation is a TROUBLEMAKER (relative to some axiology  $V$ ) if and only if there is a troublesome combination of options therein (relative to some axiology  $V$ ), i.e., if the agents can act in ways such that*

**COLLECTIVE SUBOPTIMALITY** *together they would produce a morally suboptimal outcome (relative to  $V$ ) and*

**INDIVIDUAL OPTIMALITY** *none of them could make a difference for the morally better (relative to  $V$ ) by unilaterally acting differently.*

To reveal  $P_{\exists T}$ 's structure, I need a formal notion of troublesome combinations. (In a way, the need to formalize TROUBLEMAKER-hood feels like the

optimal opportunity to also achieve a higher level of generality. But because this does not change the core of my argumentation, it is not necessary at all. Instead, as already stated in the last chapter, I continue to restrict my argumentation to minimal collective decision situations. More specifically, I continue to focus on TWO FACTORIES-like cases, i.e., COORDINATION CASES satisfying the TRIAD.)

Building on the auxiliary notion of set of 1-variants  $\Psi^{\gamma,1}$  of a combination  $\gamma \in \Psi$  (cf. Section 5.2.2), we define, given a decision situation  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle$  and a valuation function  $\text{Val} : \mathcal{O} \rightarrow \mathcal{V}$  (with an order  $<$  over  $\mathcal{V}$ ), the *set of troublesome combinations* as

$$\begin{aligned} \Psi_{\text{trouble}} := \{ \gamma \in \Psi & \mid \underbrace{\exists \gamma' \in \Psi : \text{Val}(\text{Out}_C(\gamma')) > \text{Val}(\text{Out}_C(\gamma))}_{\approx \text{COLLECTIVE SUBOPTIMALITY}} \\ & \wedge \underbrace{\nexists \gamma' \in \Psi^{\gamma,1} : \text{Val}(\text{Out}_C(\gamma')) > \text{Val}(\text{Out}_C(\gamma))}_{\approx \text{INDIVIDUAL OPTIMALITY}} \} \end{aligned}$$

This can be rewritten a bit more briefly, using the arg max notion, as

$$\begin{aligned} \Psi_{\text{trouble}} := \{ \gamma \in \Psi & \mid \underbrace{\gamma \notin \arg \max_{\gamma' \in \Psi} \text{Val}(\text{Out}_C(\gamma'))}_{\approx \text{COLLECTIVE SUBOPTIMALITY}} \\ & \wedge \underbrace{\gamma \in \arg \max_{\gamma' \in \Psi^{\gamma,1}} \text{Val}(\text{Out}_C(\gamma'))}_{\approx \text{INDIVIDUAL OPTIMALITY}} \} \end{aligned}$$

Accordingly, a decision situation  $D$  with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle$  is a TROUBLEMAKER if and only if  $\Psi_{\text{trouble}} \neq \emptyset$ , i.e., if and only if there is a troublesome combination.

Thus, we get a formal structure for  $P_{\exists T}$  that reads:

$$(P_{\exists T}) \quad \begin{aligned} \exists D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \gamma \in \Psi : \\ \gamma \notin \arg \max_{\gamma' \in \Psi} \text{Val}(\text{Out}_C(\gamma')) \wedge \gamma \in \arg \max_{\gamma' \in \Psi^{\gamma,1}} \text{Val}(\text{Out}_C(\gamma')) \end{aligned}$$

Before we turn to the semantic structure of  $P_{\text{MOCoR}}$ , it is worth stressing that the formal condition for troublesome combinations illustrates that we can frame the CHALLENGE as an optimization issue, i.e., MOAC's apparent inability to guarantee optimal solutions when used as a decision procedure. This allows us to understand TROUBLEMAKERS as challenging insofar as they allow for combinations that are local maxima. These local maxima are where MOCoR allegedly »gets stuck«. To drive this point home, consider

**Definition 6.2 (Local and global Maxima)**

*Let  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$  be a collective decision situation.*

*A combination of actions  $\Upsilon \in \Psi_\Gamma$  is a local (or relative) maximum combination if and only if  $\Upsilon$ 's consequences are a local (or relative) maximum within the evaluative profile of  $D$ , i.e., no individual deviation from that combination that would have better consequences in  $D$ .*

*A combination of actions is a global (or absolute) maximum combination if and only if  $\Upsilon$ 's consequences are a global maximum within the evaluative profile of  $D$ , i.e., there is no combination with better consequences in  $D$ .*

Trivially, every global maximum combination is a local one: If there is no combination with better consequences, then there cannot be a combination as a result of unilateral deviation that has better consequences. A bit more formally, then, we get, for some decision situation  $D$  (as above) and a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$ , that

$\Upsilon \in \Psi_\Gamma$  is a local maximum combination

if and only if

there is no  $\Upsilon' \in \Psi^{\Upsilon,1}$  with  $\text{Val}(\text{Out}_C(\Upsilon')) > \text{Val}(\text{Out}_C(\Upsilon))$

if and only if

$$\Upsilon \in \arg \max_{\Upsilon' \in \Psi^{\Upsilon,1}} \text{Val}(\text{Out}_C(\Upsilon'))$$

and

$\Upsilon \in \Psi_\Gamma$  is a global maximum

if and only if

there is no  $\Upsilon' \in \Psi_\Gamma$  with  $\text{Val}(\text{Out}_C(\Upsilon')) > \text{Val}(\text{Out}_C(\Upsilon))$

if and only if

$$\Upsilon \in \arg \max_{\Upsilon' \in \Psi_\Gamma} \text{Val}(\text{Out}_C(\Upsilon'))$$

Troublesome combinations,

thus, are exactly such combinations of actions that are local but not global maximum combinations. In other words,  $D$  is a TROUBLEMAKER if and only if it contains some genuine local maximum (i.e., a non-global local maximum).

Derek Parfit made a very similar point when he proposed to classify different kinds of COORDINATION CASES de-

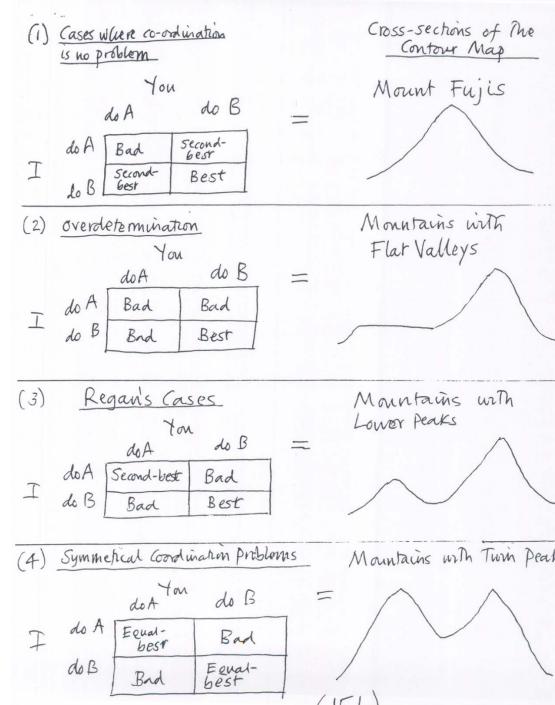


Figure 6.1: Parfit's contour maps (Parfit 1988 p.15b).

pending on the »contour maps« (cf. [ibid.](#), pp.14-15 and also Figure 6.1). In this respect, Bacharach's name for such cases as »Hi-Lo Cases« is also extremely apt (cf. Bacharach [1999](#)).

### 6.4.2 The Structure of $P_{\text{MOCoR}}$ : Ex Post!

Next, let us consider

$(P_{\text{MOCoR}})$  If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

Since this proposition obviously is meant to tell us something about all decision situations (with specific properties), we can start by getting the quantifiers right:

$$\forall D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$$

$(P_{\text{MOCoR}})$  if none of them could make a difference  
for the better by unilaterally acting differently,  
then each of them would act morally right.

Up to this point, no doubt, the naïve formalization above was on the right track. However, drawing on what we learned about  $P_{\exists T}$ , we can easily translate the antecedent in a syntactically richer way than it was offered then:

$\forall D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi :$  if  
 $(P_{\text{MOCoR}})$   $\nexists \Upsilon' \in \Psi^{\Upsilon, 1} : \text{Val}(\text{Out}_C(\Upsilon')) > \text{Val}(\text{Out}_C(\Upsilon))$   
then each of them would act morally right.

For the formalization of the consequent, we make use of two other building blocks introduced earlier. First, we need the definition of the predicate

of rightness relative to a decision situation, a context, and a moral theory (cf. page [35]), i.e.,  $T, D, C \models R\phi$ . Second, we need to apply the apparatus we introduced to capture the technique of conditionalization defined in the preceding chapter:

$$\begin{aligned} \forall D := \langle A, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ (P_{\text{MOCoR}}) \quad \nexists \Upsilon' \in \Psi^{\Upsilon, 1} : \text{Val}(\text{Out}_C(\Upsilon')) > \text{Val}(\text{Out}_C(\Upsilon)) \\ \rightarrow \forall \phi \in \Upsilon : \text{MOAC}, D \downarrow \Upsilon \ominus \phi, [\![C \oplus (\Upsilon \ominus \phi)]\!] \models R\phi \end{aligned}$$

The idea here is, basically, the CSM, i.e., that we have to assess each action of each agent against the background of all the actions of all the other agents that, together with that agent's action, make up a combination where no unilateral deviation would make a difference for the better. This enrichment of the original context indeed reduces the considered decision situations to individual decision situations that, then, can be assessed by MOAC. Since, guaranteed by the guarding of the antecedence, difference-making for the better by acting differently is impossible, the action of the respective agent under consideration, then, indeed, is assessed as right (according to MOAC).

### 6.4.3 The Structure of $P_{\text{MH}}$ : Ex Ante

We can now turn to the last remaining premise, i.e.:

( $P_{\text{MH}}$ ) If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

Recall the more precise notion of **COLLECTIVELY MAXIMIZING** we introduced in the last chapter:

**Principle 5.1 (COLLECTIVELY MAXIMIZING)** *Let  $D$  be a collective decision situation with domain  $\Psi$  and with actual context  $C$ . If  $\Upsilon \in \Psi$  consists only of right actions, then there is no (and cannot be) alternative  $\Upsilon' \in \Psi$  with better consequences than  $\Upsilon$  relative to  $C$ .*

Based on it and given the considerations in Section 6.3 as well as the already introduced formalizations above, most importantly the formalization of  $P_{\text{MOCOR}}$ , we can jump directly to the formalization of MH. First recall MH itself:

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

The straightforward formalization of MH then looks like this:

$$(P_{\text{MH}}^{\leftarrow}) \quad \forall T : \text{adeq } T \rightarrow \forall D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ \forall \phi \in \Upsilon : T, D, C \models R \phi \rightarrow \Upsilon \in \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon'))$$

Now, to get a proper formalization of  $P_{\text{MH}}$ , we only need to take the contraposition of the inner conditional. This yields (presupposing the shallow wrongness predicate from Section 2.3.4):

$$(P_{\text{MH}}) \quad \forall T : \text{adeq } T \rightarrow \forall D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \rightarrow \exists \phi \in \Upsilon : T, D, C \models W \phi$$

Note that this formalization does not involve any conditionalizing over further facts. Most importantly, we do *not* have

$$\begin{aligned} \forall T : \text{adeq } T \rightarrow \forall D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ (P_{\text{MH}}^{\rightarrow}) \quad \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \\ \rightarrow \exists \phi \in \Upsilon : T, D \downarrow_{\Upsilon \ominus \phi}, \llbracket C \oplus (\Upsilon \ominus \phi) \rrbracket \vDash W \phi \end{aligned}$$

Now that we have plausible and defensible formalizations of the three premises, we can turn to the overall ARGUMENT.

#### 6.4.4 Putting Things Together

If we put the three formalizations together (and add the very trivial conclusion), we get a reasonable structure of the ARGUMENT:

**Argument:** The ARGUMENT (formally)

$$\begin{aligned} P_{\exists T}: \quad \exists D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi : \\ \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \wedge \Upsilon \in \arg \max_{\Upsilon' \in \Psi^{\Upsilon, 1}} \text{Val}(\text{Out}_C(\Upsilon')) \end{aligned}$$

$$\begin{aligned} P_{\text{MOCoR}}: \quad & \forall D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ & \nexists \Upsilon' \in \Psi^{\Upsilon, 1} : \text{Val}(\text{Out}_C(\Upsilon')) > \text{Val}(\text{Out}_C(\Upsilon)) \\ & \rightarrow \forall \phi \in \Upsilon : \text{MOAC}, D \downarrow_{\Upsilon \ominus \phi}, \llbracket C \oplus (\Upsilon \ominus \phi) \rrbracket \vDash R \phi \end{aligned}$$

$$\begin{aligned} P_{\text{MH}}: \quad & \forall T : \text{Adeq } T \rightarrow \forall D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \forall \Upsilon \in \Psi : \\ & \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \rightarrow \exists \phi \in \Upsilon : T, D, C \vDash W \phi \end{aligned}$$


---

$$C_{\neg \text{Adeq}}: \quad \neg \text{Adeq MOAC}$$

It is not hard to see that the conclusion can not be derived from the premises. The ARGUMENT is invalid. This, somewhat surprisingly, does not mean that we are done with the CHALLENGE as CHALLENGE<sub>int</sub>.

This sobering assertion certainly needs a bit of explanation. Thus, formally, **CHALLENGE<sub>int</sub>** is solved, and we could immediately turn to the **No-DIFFERENCE CHALLENGE** (compare according to Figure 6.2 from Chapter 4).

Unfortunately, on closer inspection, the matter is much more complicated since this victory implies remaining guilt of MOAC: It remains to show that, in every **TROUBLEMAKER**, if a troublesome combination were realized, at least one action was not right. To see why this remains to be shown, we simply<sup>121</sup> infer from  $P_{\exists T}$  and  $P_{MH}$  that

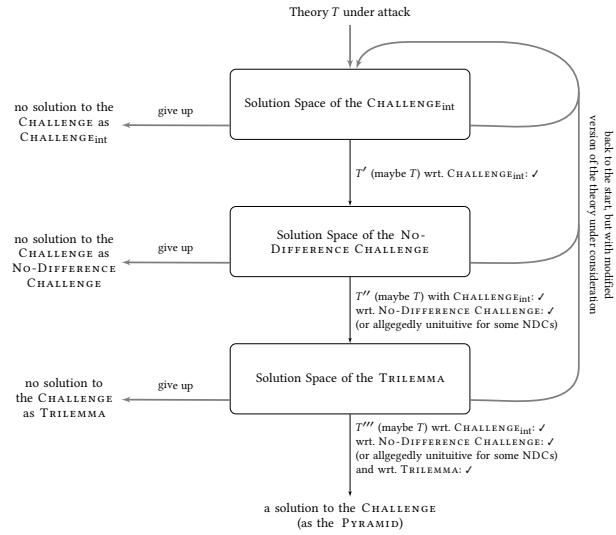


Figure 6.2: The solution space of the **CHALLENGE** as the **PYRAMID**.

$$\exists D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi :$$

$$(\text{Gaps!}) \quad (\text{Adeq MOAC} \rightarrow \exists \phi \in \Upsilon : \text{MOAC}, D, C \models W \phi)$$

$$\wedge \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \wedge \Upsilon \in \arg \max_{\Upsilon' \in \Psi^{\Upsilon, 1}} \text{Val}(\text{Out}_C(\Upsilon'))$$

Translated back, this simply means that there are **TROUBLEMAKERS** and **MOAC**, in order to be not adequate according to its own standards, must either identify wrongdoing in those situations.

<sup>121</sup>The inference has this form: We can rewrite an instantiation of  $P_{MH}$  with **MOAC** to a formula of the structure  $\neg Fa \vee \forall x : (Gx \rightarrow Hx)$  which we can rewrite to  $\forall x : \neg Fa \vee (Gx \rightarrow Hx)$  which, in turn, we can rewrite to  $\forall x : (Fa \wedge Gx) \rightarrow Hx$ . Since we can rewrite  $P_{\exists T}$  as a formula with the structure  $\exists x : Gx \wedge Ix$ , we can infer from there two that  $\exists x : (Fa \rightarrow Hx) \wedge Gx \wedge Ix$ .

In the remainder of this chapter, I will argue that MOAC actually cannot fulfill this condition and that MOAC's structural inability to fulfill (Gaps!) substantially leads the way to a new collective challenge: The REAL CHALLENGE.

## 6.5 When The Villain Finally Enters The Stage: The REAL CHALLENGE

At this point, we have realized that MOAC does not assess all actions of a troublesome combination as right when this combination is performed. In this respect, there is no direct violation of MH. However, there is a requirement implied by MH that MOAC *does not meet*. The remaining challenge, thus, is to show that at least one of the agents acted wrongly when a troublesome combination of actions is performed in a TROUBLEMAKER. However, recall that we originally introduced MH in a positive formulation, i.e.:

**Criterion 1.2 (MORAL HARMONY (MH, tentative))** *A moral theory is adequate only if it is true that if all agents act rightly (i.e., according to this theory), then they are guaranteed to produce the morally best outcome (i.e., according to this theory) they could together bring about.*

However,  $P_{\text{MH}}$  is using MH in a negative version with a contrapositive of the original condition:

**Criterion 1.3 (MORAL HARMONY (MH, tentative, contraposition))**  
*If a moral theory is adequate, then, if the agents in a collective decision situation produce a morally suboptimal outcome (i.e., according to this theory), (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).*

However, as indicated already in the introduction,<sup>122</sup> there is a non-innocent translational element in this paraphrase. Assume, instead, we would go with

**Criterion 6.1 (MORAL HARMONY (MH, tent., contrapos., alternative))**

*If a moral theory is adequate, then if the agents in a collective decision situation were to act in ways such that together they would produce a morally suboptimal outcome, then (necessarily) at least one of the agents acted not rightly (i.e., according to this theory).*

If we had used this formulation as the basis for the  $P_{\text{MH}}$  and thus the ARGUMENT, we would have ended up not with (Gaps!) above, but with

$$\begin{aligned} \exists D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle : \exists \Upsilon \in \Psi : \\ (\text{Gaps?}) \quad \wedge (\text{Adeq MOAC} \rightarrow \exists \phi \in \Upsilon : \text{MOAC}, D, C \not\models R \phi) \\ \wedge \Upsilon \notin \arg \max_{\Upsilon' \in \Psi} \text{Val}(\text{Out}_C(\Upsilon')) \wedge \Upsilon \in \arg \max_{\Upsilon' \in \Psi^{\Upsilon, 1}} \text{Val}(\text{Out}_C(\Upsilon')) \end{aligned}$$

(Gaps?) would actually be trivially fulfilled, given that one of the results of this investigation is that there is no right action whatsoever for any agent in such cases. (It should be noted that these considerations are the reason we considered the two wrongness predicates for MOAC at the end of chapter Section 2.3.4 (cf. page 54). A version of (Gaps!) based on the shallow consequentialist wrongness predicate  $W_s$  would immediately collapse into (Gaps?). Only the deep consequentialist wrongness predicate  $W_d$  allows us, at this point, to semantically distinguish the two meanings we are trying to capture.) If there is no unconditionally right action for any of the agents, obviously no agent performs a right action in a TROUBLEMAKER no matter what they

---

<sup>122</sup>Cf. Footnote 8 and Footnote 73.

do, i.e., even if an agent contributes to realizing a troublesome combination, they do no right.

There are at least three reasons why the champions of MOAC should accept the stronger and thus threatening reading and thus embrace (Gaps!) instead of (Gaps). The first reason is strategic, the second is reconstructive, and the third is systematic in character.

The strategic argument is that it is blatantly dangerous not to accept the stronger reading. Even if one sees no reason for the stronger reading (and there are two such reasons immediately following), one had better face a threatening challenge unless one can make a very strong argument for the weak reading (and I know of none). In other words, even if they would turn their back to CHALLENGE<sub>int</sub> now, they should not be able to sleep well. As soon as someone comes up with an argument for the stronger reading, they have to face that challenge either way. One would leave an open flank that might expose a significant vulnerability.

The reconstructive argument is that the stronger reading actually corresponds to some formulations in the literature and seems to do more justice to the general idea of PMH. With respect to the first point, we can refer primarily to Pinkert's ON-THE-HOOK (cf. Footnote 73 and more generally page 122), we recall:

**Principle 3.8 (ON-THE-HOOK)** *In any collection of agents who together gratuitously fail to bring about collectively optimal outcomes, there must be some relevant morally objectionable facts about some of the agents.*

Pinkert's formulation is explicitly meant to be a contraposition of a more general principle (cf. Pinkert 2015, p.976, my italics):

I contend that for this reason, On-the-hook should not be understood as a specifically Consequentialist position. Instead, it should be understood as the *contraposition of a second-order claim about morality in general* and, hence, as a desideratum for any moral principle.

Clearly, the call for the existence »morally objectionable facts« is closer to the idea of looking for wrongdoings. If Pinkert advocated for a version that calls for the absence of »morally commendable facts«, this would correspond to Criterion 6.1's formulation of missing rightdoings.

But I do not just want to appeal to Pinkert's authority here. Instead, Pinkert seems just right about this formulation. The general idea inherent in PMH is forward-looking<sup>123</sup> PMH is about the claim that morality has the function of pointing out the way to the optimal. It is not so much concerned with the idea that morality is merely not pointing the way to the suboptimal. In this respect, according to the PMH, it should be specifically wrong to perform actions that exclude achieving the optimum, as would be the case for many TROUBLEMAKERS for those actions that constitute problematic actions.

Thirdly, there are systematic considerations that should let the consequentialist search for a solution, which allows him to give non-conditional assessments for the agents' choices within TROUBLEMAKERS. After all, a closer look at (Gaps!) reveals a deeper problem that doesn't have much to do with TROUBLEMAKERS at all. Instead, the challenge ultimately lies in the fact that MOAC simply cannot provide a satisfactory answer for many overtly

---

<sup>123</sup>We might refer to the citations of Feldman and Portmore in Section 6.3 above. But also, we might remember Broad's positive variant of the FALSE UNIVERSALIZATION from Section 3.5.1 which was meant to equip agents pondering on their options with a forward-looking principle that allows them to make the right decisions.

morally charged decisions that agents may face. MOAC fails to provide non-conditional assessments for all non-separable collective decision situations. This is too little for an objective moral theory that naturally would long to have DEONTIC COMPLETENESS, not in a purely formal but a substantive way. By this, I mean that MOAC should not retreat to the GENUINE KIND VIEW and then pretend that a property like DEONTIC COMPLETENESS (or even WEAK DEONTIC COMPLETENESS) is formally attributable to MOAC because the situations in which the agents are in such non-separable collective decision situations are not individual decision situations at all. After all, these situations are such that agents must choose from a set of options. To withdraw as MOAC in such situations to the position that one can only give ›it depends on what the others do‹-answers, possibly with the (lame!) excuse that the agents are not in a *real* (i.e., individual) decision situation at all, seems not only theoretically unsatisfying but becomes close to committing a no true Scotsman fallacy.

Thus, if we follow these reasons, we are ultimately dealing with a weakness of MOAC that extends far beyond TROUBLEMAKERS and that is not directly connected to PMH even at its core. In light of METHODOLOGICAL INDETERMINISM, MOAC is simply incapable of working out which actions are right and wrong in multi-agent situations where the outcomes of the agents' actions are sufficiently interdependent. It just cannot identify *the consequences* of the individual agent's actions. MOAC can only assess retrospectively or under the assumption of already determined future actions. Both ways lead to disaster.

Since we assume METHODOLOGICAL INDETERMINISM, we can conclude: MOAC fails to live up to its desideratum of a DEONTIC COMPLETE-

NESS (cf. Section 4.3.2). Instead, a closer look reveals a world of choices that takes place in moral gaps. Only when actions have filled these Gaps can it be said from the perspective of MOAC whether these actions were right or wrong. This challenge, the REAL CHALLENGE, I would like to suggest can be grasped argumentatively as follows in terms of the REAL ARGUMENT:

**Argument:** The REAL ARGUMENT

$P_{\exists NS}$ : There are non-separable collective decision situations, i.e., collective decision situations in which, relative to the actual and normatively sufficiently complete context, it is not defined for at least one agent with respect to at least two of their options how they are to be ranked with respect to the moral quality of their actual consequences.

$P_{\exists GAPS}$ : If there are such collective action situations, then there is a widespread class of decisions to which MOAC has nothing illuminating to contribute; that is, there are a myriad of systematic deontic gaps.

$P_{\text{No GAPS}!}$ : If a moral theory is adequate, then there are no widespread class of decisions to which the theory has nothing illuminating to contribute; that is, there are no systematic deontic gaps.

---

$C_{\neg \text{Adeq}}$ : MOAC is not adequate.

The CHALLENGE, then, ultimately relates to the REAL CHALLENGE as a symptom that has its cause in a particular tactical movement – the CONSEQUENTIALIST STANDARD MOVE (CSM) – that MOAC has made to

manage the REAL CHALLENGE: It would be unbearable for MOAC to be unable to make genuine unconditional assessments in a widespread class of collective decision situations. However, they apparently can only give *conditional* assessments in all interesting collective decision situations. To avoid this, the champions of MOAC are willing to allow the consideration of facts about the actual actions of the respective other agents to resolve the conditional assessments. The champions of MOAC, thus, *implicitly assume* that the original descriptions of the cases were incomplete, i.e., they implicitly reject  $P_{\exists NS}$ . Thus, they pretend there were normatively relevant facts missing – namely, facts about how the agents actually act. This move, the addition of factual actions, however, allows us to find situations – namely, TROUBLEMAKERS – in which the dependencies are so unfavorable that one finds combinations of actions in which the actions have to be assessed as being right, even though the involved agents collectively bring about sub-optimal and unbearable overall consequences. This is the CHALLENGE: a consequence of the REAL CHALLENGE together with a quick-and-dirty fix, i.e., the CSM.

At this point, it is worthwhile to look back and consider Regan's impossibility proof from this point of view (cf. Section 3.5.2). So, here is Regan again with his step-by-step application of the CSM. We begin by recalling Regan's toy example (Regan 1980, p. 18), structurally equivalent to Two FACTORIES:

**Case 3.7 (WHIFF AND POOF)** *Suppose that there are only two agents in the moral universe, called Whiff and Poof. Each has a button in front of him which he can push or not. If both Whiff and Poof push their buttons, the consequences will be such that the overall state of the world has a value of ten units. If neither Whiff nor Poof pushes his button, the consequences will be such that the*

*overall state of the world has a value of 6 units. Finally, if one and only one of the pair pushes his button (and it does not matter who pushes and who does not), the consequences will be such that the overall state of the world has a value of 0 (zero) units. Neither agent, we assume, is in a position to influence the other's choice.*

This case can be represented in the following normal form:

		Poof	
		not-push	push
Whiff	not-push	6	0
	push	0	10

Note that this description contains no assumptions regarding what Whiff and Poof actually do or will do. He starts from the case as given here.

Next, we return to what Regan called a »precise necessary condition for exclusive act-orientation« (Regan [1980] p.114) that he calls »the partial definition«. Recall that the property of exclusive act-orientation is the core piece of Regan's result:

Any exclusively act-oriented theory must, in this example, on any assumption about Poof's (Whiff's) behavior, identify some non-empty subset of the set of acts comprising »pushing« and »not-pushing« such that Whiff (Poof) satisfies the theory if and only if he does some act from that subset.

Notice that, by that partial definition, every exclusively act-oriented theory is meant to deliver a *non-empty* subset of the option space of an agent. In a sense, this is a built-in commitment to  $P_{\exists GAPS}$  for such theories. Regan gives an argument for this commitment (*ibid.*, p.115):

If an exclusively act-oriented theory selected the empty subset on any assumption about Poof's (Whiff's) behaviour, then it would direct Whiff (Poof) to do the impossible. (Selecting the empty subset is not the same as directing the agent not to push. »Not-pushing« is an act for our purposes. Selecting the empty subset is directing the agent to neither push nor not-push, which he cannot do.)

In other words, Regan claims that if we were to allow empty sets of right actions, this would mean violating some version of the infamous »ought« implies »can«. If the relevant set was empty relative to some decision situation and an agent, MOAC would still require that agent to perform one of the actions within that set. Thus, the agent ought to perform *one of none* actions – which is logically impossible and, hence, they cannot do. This is why Regan built RESOLVABILITY in his partial definition.

Next recall Regan unfolding his central argument ([ibid.](#) p. 115):

Suppose there is an adaptable theory  $T$  which satisfies the partial definition. Suppose further that Poof does not push. Since  $T$  satisfies the partial definition, there is some non-empty subset of the set of acts »pushing« and »not-pushing« such that Whiff satisfies  $T$  (while Poof does not push) if and only if he does an act from that subset. Call the subset  $S$ . We can deduce what  $S$  must be from the assumptions we have made about  $T$ . We know that Whiff satisfies  $T$  if and only if he does an act from  $S$ . So, if Whiff does an act from  $S$ , he satisfies  $T$ . Since  $T$  adaptable,  $T$  [embraces MOCOR]. That means that any agent who satisfies  $T$  produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, which include Poof's not-pushing, he must not-push. Therefore, if Whiff satisfies  $T$ , he not-pushes. Remembering what we have already established, that if Whiff does an act from  $S$ , he satisfies  $T$ , we can conclude that if Whiff does an act from  $S$ , he not-pushes.

But remember also that  $S$  is non-empty. The only non-empty set such that if Whiff does an act from that set he not-pushes is of course the set consisting of the act »not-pushing«. Therefore  $S$  consists of the act »not-pushing«. In sum, if Poof does not push, then Whiff satisfies  $T$  if and only if he (Whiff) not-pushes also.

According to my analysis, things go wrong instantly, right at the second sentence, where Regan assumes that Poof does not push, adding this fact to Whiffs »circumstances«. Actually, given Regan's (MOCOR-like) explications of the act-consequentialist criterion of rightness, Regan had to smuggle this into his proof. Consider a version without that:

Suppose there is an adaptable theory  $T$  which satisfies the partial definition. Since  $T$  satisfies the partial definition, there is some non-empty subset of the set of acts »pushing« and »not-pushing« such that Whiff satisfies  $T$  (while Poof does not push) if and only if he does an act from that subset. Call the subset  $S$ . We can deduce what  $S$  must be from the assumptions we have made about  $T$ . We know that Whiff satisfies  $T$  if and only if he does an act from  $S$ . So, if Whiff does an act from  $S$ , he satisfies  $T$ . Since  $T$  adaptable,  $T$  [embraces MOCOR ]. That means that any agent who satisfies  $T$  produces the best possible consequences in his circumstances. If Whiff produces best possible consequences in his circumstances, he must ????. Therefore, if Whiff satisfies  $T$ , he ??. Remembering what we have already established, that if Whiff does an act from  $S$ , he satisfies  $T$ , we can conclude that if Whiff does an act from  $S$ , he ??. But remember also that  $S$  is non-empty. The only non-empty set such that if Whiff does an act from that set he ??? is of course the set consisting of the act ??. Therefore  $S$  consists of the act ??. In sum, if Poof does ???, then Whiff satisfies  $T$  if and only if he (Whiff) ??? also.

The problem here is that Regan's partial definition above ensures us that there is a non-empty set  $S$  of right actions according to MOAC. However, apparently, there is *no* such set  $S$  given WHIFF AND POOF. Only if we add a fact about what the other does, there comes such set into existence, namely one given WHIFF AND POOF plus that fact. But then, at least one of them cannot satisfy  $T$  in any meaningful way: Whoever acts first has performed an act that was not in >his< set  $S$ , because for him the set of right actions was empty. Further, note that Regan was completely aware that he had to add such an additional fact. Quite early in his book he states (Regan 1980, p. 18):

Now, if we ask what AU [i.e., MOAC] directs Whiff to do, we find that we cannot say. If Poof pushes, then AU directs Whiff to push. If Poof does not push, then AU directs Whiff not to push. Until we specify how Poof behaves, AU gives Whiff no clear direction. The same is true, *mutatis mutandis*, of Poof.

But instead of turning that observation into an argument against MOAC, Regan focussed on showing that MOAC fails wrt. to PMH.

In light of all these, I think we should better understand Regan's overall argument like this: If a moral theory wants to deserve to be called exclusively action-oriented, then it must make the moral status of actions exclusively a function of those actions. Moreover, a moral theory should be generally applicable in the sense that it should be applicable to the choices of agents, at least by default. As mentioned above, it seems deeply unsatisfactory to tell Ann and Ben in Two FACTORIES that MOAC has nothing to say about their choices because they are in a collective decision situation and, actually, do not face >true< individual decision situations. Further, a moral theory must never demand the impossible of agents. Specifically, there must always be an action that, if performed, satisfies the theory (. TROUBLEMAKERS show, however,

champions of MOAC (and, in fact, of any exclusively action-oriented moral theory) face a theoretical *trilemma*: either they give up general applicability, i.e., DEONTIC COMPLETENESS, because they have nothing to say in such cases; or they demand the impossible, namely requiring the agents to perform a right action without being able to name one, which implies to give up NO MORAL DILEMMAS; or they add facts about what the other agents do or will do to such cases, violating METHODOLOGICAL INDETERMINISM. However, these theories cannot satisfy PMH in the latter case. Since this is also not an acceptable option for camp MOAC, i.e., this is the CHALLENGE as the CHALLENGE<sub>int</sub>, MOAC is irredeemably lost.

The main point I made in this chapter is then this: the CHALLENGE as CHALLENGE<sub>int</sub>, given closer inspection, does not *really* work as planned because there are no strict violations of the PMH. As the agents, according to this quickly painted picture, act first and are assessed in retrospect, they can neither follow morality nor satisfy moral theories when acting, nor can they fail to do so. The challenge, thus, lies deeper: the core idea of PMH just runs empty – due to the deontic Gaps that were never filled meaningfully and substantially but were only argumentatively hastily patched over.

In light of all this, it becomes obvious what camp MOAC ultimately needs: A way to resolve the REAL CHALLENGE *without* running into any version of the CHALLENGE, i.e., in a way that allows MOAC to fill the deontic gaps, to give unconditional assessments that are, actually, in line with (consequentialist) basic intuitions. This is what remains to be done and what I strive to do in the remainder of this thesis.

# **Chapter 7**

## **Of New Consequences**

This chapter aims to illustrate and spell out my central approach, which I will call the **INTERMEDIATE OUTCOMES APPROACH** (or just: the **APPROACH**). I discuss how it can help to actually fill the identified deontic gaps by exploiting *newly discovered consequences* of arbitrary actions in collective decision situations. The result is a *multi-agent version of consequentialism* equipped with the principled capability to provide non-conditional assessments for interesting collective decision situations, including **TROUBLEMAKERS**. Along the way, we will find a new, *unified representation of collective decision situations* that fits well with the **APPROACH** and allows us to visualize both the **CHALLENGE** and the **REAL CHALLENGE**. This will make us realize that there are actually several ways of exploiting the newly discovered consequences. I will call these *multi-agent amendments*, and I will present a selection of amendments that I take to contain the most interesting and important candidates. At the end of this chapter, there remains only one decision left: which amendment MOAC should embrace – and why. This, then, will be the topic of the next and last substantial chapter.

## 7.1 New Grounds for Consequentialism

At first glance, the matter might look hopeless. To solve the REAL CHALLENGE, MOAC must be able to give non-conditional assessments of all the agents' options (or at least of their ›right‹ options) in interesting collective decision situations. But to not give up its act-consequentialist roots, it must not resort to anything other than the consequences of these options. Yet interesting collective action situations – and thus TROUBLEMAKERS – are characterized precisely by the fact that the consequences of actions depend essentially on what the other agents do. Hence, COMPOSITIONALISM was also, to all appearances, a hopeless position, while the GENUINE KIND VIEW prevailed. We recall:

**Claim 5.1 (COMPOSITIONALISM)** *All collective decision situations can be reduced to individual decision situations (plus some structure).*

**Claim 5.2 (GENUINE KIND VIEW)** *Some collective decision situations cannot be reduced to individual decision situations (plus some structure).*

How can the consequentialist extricate himself from this apparently hopeless situation?

The answer is: He must defend COMPOSITIONALISM and thus deny the existence of apparently non-reducible and, hence, of allegedly interesting collective decision situations. For this, he ›merely‹ has to show how apparently interesting collective decision situations could be decomposed into individual decision situations. What MOAC lacks for this are, obviously, consequences that can be directly assigned to individual actions without conditionalizing over other agents' actions. But from where to take such consequences?

The answer is so evident that one can only say that MOAC did not see the forest for the trees. For, of course, we have long since seen the relevant consequences and have even listed them quite explicitly several times in the preceding chapters. In the following, I will mark the forest and then defend that this is not a ›lunatic insight‹.

### 7.1.1 »Like Scales Fell From His Eyes...«

At this point, we are in search of overlooked consequences. This first raises the question of what we are actually looking for when we search for consequences. The following rough characterization will probably suffice to give us a reasonably reliable criterion for what counts as a consequence of some given action  $\phi$ : We are looking for something that would be the case if  $\phi$  were performed and would be absent if, instead of  $\phi$ , a specific other action was performed. This should remind us of Jackson's

**Principle 3.1 (Difference Principle)** *The morality of an action depends on the difference it makes; [i.e.,] it depends on the relationship between what would be the case were the act performed and what would be the case were the act not performed.*

So, let's return to our running example to look for such differences. Recall:

**Case 3.1 (Two Factories)** *Ann and Ben each own a factory near the same river. Both can produce either cleanly, or cheaply and thereby pollute the river. The local market is highly competitive. Thus, a factory that produces cleanly would become noncompetitive if the other factory pollutes. The local social system is underdeveloped, and the economic situation is terrible. Hence, if*

*a factory closes, this will cause significant unemployment and social hardship. If at least one factory produces cheaply, the resulting pollution will eventually destroy the local ecosystem and erode the livelihood of a village downstream. However, any additional polluter would not make the situation worse in this regard. Ann and Ben decide and act independently in the sense that neither of them can coordinate with the other, nor can any agent observe the actions of the other. Thus, whatever each agent does, they would do it regardless of what the other does.*

The straightforward representation of the TROUBLEMAKER in terms of a normal form was given by this tabular:

		Ben	
		pollute	produce cleanly
		pollute	second-worst
Ann			worst
	produce cleanly	worst	best

The essential question now is this: What difference does it make if Ann produces cleanly? The answer is: the decision situation that remains for Ben! That insight, in a way, has been the whole idea behind conditionalization. Suppose

**Fact 7.1** *Ann produces cleanly.*

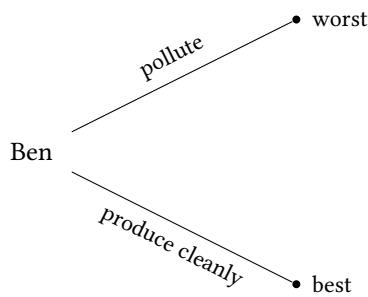
then we have ensured ourselves that

(44) It is right for Ben to produce cleanly.

and

(45) It is wrong for Ben to pollute.

are true relative to  $\llbracket \text{Two FACTORIES} \oplus \text{Fact 7.1} \rrbracket$ . We have already established that Fact 7.1 allows us to reduce Two FACTORIES to this individual decision situation of Ben:



However, what would be the case if Ann would do otherwise? Let us entertain:

**Fact 7.2** *Ann pollutes.*

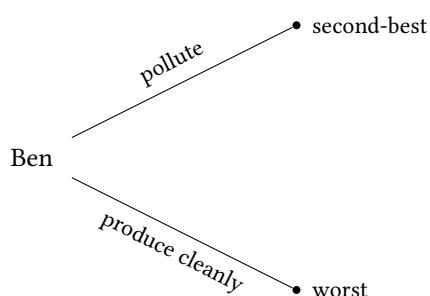
Relative to  $\llbracket \text{Two FACTORIES} \oplus \text{Fact 7.2} \rrbracket$  it is true both

(46) It is right for Ben to pollute.

and

(47) It is wrong for Ben to produce cleanly.

Further, Fact 7.2 allows us to reduce Two FACTORIES to this individual decision situation of Ben:



Thus, if we are looking for differences Ann's actions make, then these two remaining individual decision situations of Ben, the results of the reduction of **TWO FACTORIES** relative to Ann's respective action, are actually pretty plausible candidates. The other way around, of course, applies to Ben as well: if he acts first, this brings about one of two possible remaining individual decision situations for Ann. Looking at the normal form of the case, we can express this insight like this: If Ann acts first, the consequence of her action corresponds to the line associated with her action; if Ben acts first, the consequence of his action corresponds to one of the two columns.

This insight might very well help us solve both the **REAL CHALLENGE** and the **CHALLENGE**. If we accept these outcomes as proper outcomes *and* if we find a way to utilize these outcomes appropriately, then we might fill the gaps in a way that does not reiterate the **CHALLENGE**. In the remainder of this thesis, I turn to why we should accept these outcomes and how to utilize them appropriately. But for now, let's concentrate on the general idea first.

Two questions may arise: first, how should we assess the later acting agents' actions, and second, what should we say in the rather esoteric edge case of simultaneous action? The answer to the first question is particularly straightforward in the two-agent case. Whoever acts second is in their remaining individual decision situation brought about by the first acting agent. To assess this situation is not a challenge, but the daily business of MOAC and MOCoR. In cases with more agents, we can just reiterate the procedure: The result of the action of the second-agent action is a further reduced decision situation. We can recursively apply the same procedure until, finally, we end in a classical individual decision situation of the last-acting agent.

Regarding the synchronous case, things also seem pretty straightforward: Since the actual context of their decision cannot contain any fact of what the other agent does up to the moment of acting (for methodological reasons, recall **METHODOLOGICAL INDETERMINISM**), both bring about the individual decision situation of the other agent. Even though the other agent's decision situation is immediately resolved in this case of synchronous action, this does not affect the assessment in the original context.

Call this approach, i.e., the consideration of the remaining decision situation of the other agent (or agents) as the consequence of an agent's actions within a collective decision situation, **APPROACH**. It means breaking down the one-step approach underlying our current understanding of collective decision situations into a step-by-step approach: Instead of just assigning *final outcomes* to combinations of actions, we start with a first action that reduces the collective decision situation to some intermediate outcome that itself is a decision situation. If more than two agents are involved, a sequence of further action-by-action reductions finally leads to an individual decision situation before the last action leads to a final outcome.<sup>124</sup>

It instantly follows that we can maintain **COMPOSITIONALISM**. Recall

**Claim 5.1 (COMPOSITIONALISM)** *All collective decision situations can be reduced to individual decision situations (plus some structure).*

Here is how we can decompose any arbitrary collective decision situation into individual ones. Let  $D$  be a collective decision situation with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$ . Then, we can decompose  $D$  in  $n := |\mathcal{A}|$  individual decision situations  $D_{A_1}, \dots, D_{A_n}$ , one for each agent, with the structure

$$D_{A_i} := \langle A_i, \Phi_{A_i}, \text{Out}_C : \Phi_{A_i} \rightarrow \mathcal{O}_D^{A_i} \rangle$$

---

<sup>124</sup>Note that we restricted this investigation to finite collective decision situations.

where

$$\mathcal{O}_D^{A_i} := \{ D_{\downarrow \phi} \mid \phi \in \Phi_{A_i} \}.$$

Note how all these decision situations are structurally interconnected because the outcomes of each of them are reduced versions of the original collective decision situation  $D$ , involving all the other agents and their remaining choices.

Let's call the set  $I_D := \{ D_{A_i} \mid A_i \in \mathcal{A}_D \}$   *$D$ 's decomposition*.

So it turns out that the so attractive and formerly apparently hopeless claim **COMPOSITIONALISM** prevails over **GENUINE KIND VIEW** after all. Recall that **COMPOSITIONALISM** is so attractive for MOAC because, if true, it would unlock the universal applicability of MOCOR in the collective domain – given that the newly discovered consequences can be ranked as required by MOCOR. This should make the champions of MOAC optimistic.

This brings us back to the central task that remains to be done. For **APPROACH** to be ultimately useful for MOAC, it must be decided how the newly discovered consequence, i.e., the reduced decision situations, are to be integrated into MOAC's overall framework. This comes down to deciding how these decision situations are to be evaluated (or at least ranked) morally and, thus, to an axiological question in the broad sense.<sup>[125]</sup> However, this question of the appropriate moral assessment of decision situations as consequences is anything but trivial as different amendments have to be considered and, somehow, compared.

In the remainder of this chapter, however, I would first like to emphasize that the approach I propose is anything but far-fetched and is much more wonderfully embedded in more general consequentialist thought. Then, I will look again at PMH for reflections on how PMH might (not) help us

---

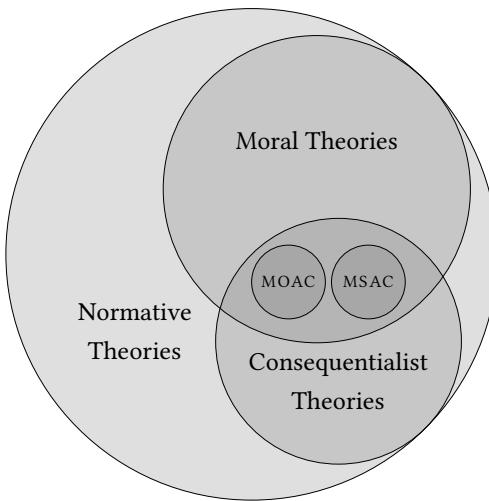
<sup>[125]</sup>Recall Section 2.3.3 on 46.

choose an appropriate answer to the question of moral assessment of decision situations. Finally, I will collect some candidates for amendments. I leave the assessment and comparison of these amendments – and the question of how to compare them – to the next chapter.

### 7.1.2 Exotic and Esoteric? Or Old Wine in New Bottles?

MOAC is a consequentialist *moral* theory. However, there are other, non-moral kinds of consequentialist theories for which the idea of decision situations being consequences is prevalent (for an overview of the different types of normative theories at play here, see

Figure 7.1). For instance, classical decision theory, understood as a theory of instrumental rationality, has long recognized a nuanced pic-



**Figure 7.1:** How normative, moral, and consequentialist theories relate to each other. Standard decision theory, as a theory of instrumental rationality, is a non-moral consequentialist theory.

ture of kinds of outcomes. Most importantly, the von Neumann–Morgenstern utility theorem, one of the fundamental theorems of utility theory, lets agents choose between *lotteries*, scenarios with uncertain outcomes, some of which might be lotteries themselves.<sup>126</sup> The utility of these lotteries is then thought of as a function of the possible *final* outcomes and the probabilities that they obtain, typically their expected utility. Insofar lotteries can be considered a

<sup>126</sup>If the uncertainty can be expressed in terms of probabilities such decisions are typically described as decisions under *risk*. This is distinct from decisions under *uncertainty*, where outcomes lack specific probabilities.

kind of *intermediate* outcomes.

Game theory, with its intrinsic focus on multi-agent dynamics of decisions and actions, entertains the idea of decision situations as (intermediate) outcomes even further. As agents maneuver within games, the choices of one agent typically are equivalent to the selections of sub-games for other agents. From the perspective of game theory, an extensive form, as seen many times in this thesis already, is a tree (or at least a direct acyclic graph) representing the (remaining) game, where transitions correspond to the agents' actions. Every action corresponds to the selection of a sub-tree, and the consequences of the actions correspond to the remaining game and, thus, to the decision the following agent has to make.

The readiness to accept decision situations as outcomes is not restricted to the consequentialist domain of instrumental rationality, though. Many subjective consequentialist theories like MSAC are rooted in expected utility theory and, more generally, decision theory. Recall

**Principle 3.9 (MSCoR (prototypical))** *It is right to perform a certain action if and only if there is no alternative with expectedly better consequences.*

Such theories can easily consider scenarios where actions' immediate consequences are lotteries. Recall the original version taken from Frank Jackson (Jackson 1991, p. 462):

**Case 3.8 (THE DRUG)** *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of the other two*

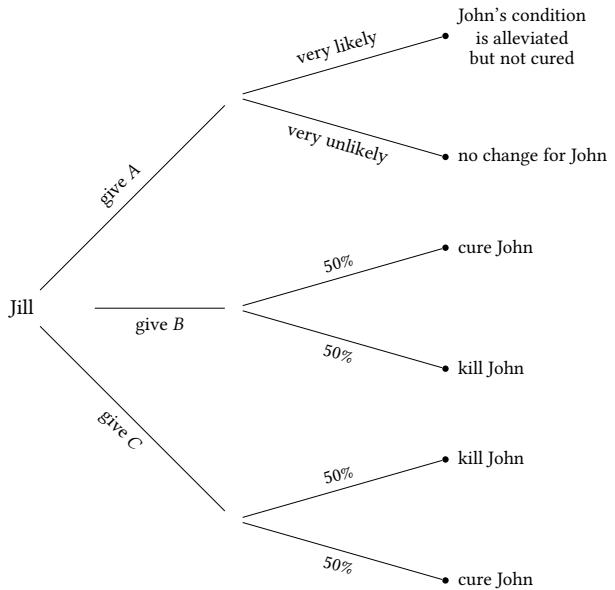
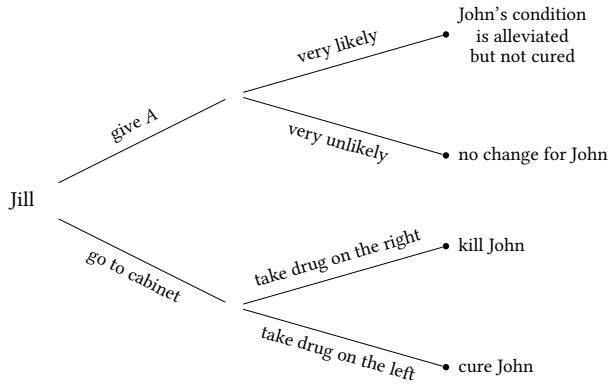


Figure 7.2: The extensive forms of THE DRUG.

*drugs, either B or C, will completely cure the skin condition; the other though will kill John, and there is no way that she can tell which of the two is the perfect cure and which is the killer drug.*

In Chapter 3 we represented THE DRUG as a decision tree (cf. Figure 7.2), involving three lotteries. But we can easily modify the case into a sequence of two subsequent decisions without introducing any riddles for camp MSAC such that the resulting case is basically equivalent to the original. Consider

**Case 7.1 (THE DRUG (II))** *Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it, drug B will completely cure the skin condition, and drug C will kill John. Drug A is on the tray directly in front of Jill, while drugs B and C are in the cabinet in the next room. However, the labels of these two drugs in the cabinet are illegible due to*



**Figure 7.3:** The extensive forms of THE DRUG (II).

*the passage of time, and there is no way for Jill to figure out which one is which. As a matter of fact, but unbeknown to Jill, the drug standing on the left is drug B, and the drug on the right is drug C.*

If Jill decides against drug A, she decides to go to the cabinet, where she faces another decision. We may assume that if Jill had to pick between the two, she would select randomly with a probability of 50%. It is easy to see that Jill's new decision is very similar and open to the same analysis and assessment as the original THE DRUG case (cf. Figure 7.3). But in THE DRUG (II), Jill can either decide on an action with a lottery outcome (giving the suboptimal drug A) or bring herself into another decision situation. This should not bother the subjective consequentialist at all. And lotteries and decision situations are not so different after all, as decision theory demonstrates – and Kagan's discussion of the CHALLENGE demonstrates, for instance, that subjective consequentialists also see it this way (cf. Kagan 2011 and Section 4.4.1).

For object consequentialism, however, it seems to remain at least an unusual conceptualization to think of agents' actions as lotteries – or, more specifically, selections of other agents' decision situations. Historically, objective consequentialism has focused primarily on the singular choices of individual agents *without uncertainty*. There was not much space for exogenous

variables. And if they were considered, then typically not so much as being dynamic elements of change, but rather as static and pre-determined. Though providing clarity in specific scenarios, such an approach inadvertently narrows the consequentialist lens, omitting the more intricate tapestry.

This is certainly not to say that *no one* at camp MOAC has thought about decisions with genuine uncertainty. That would be more than surprising given the challenge that Prior has formulated (cf. A. N. Prior and Raphael [1956; see above]). Indeed, quite some authors have thought about the indeterminacy or under-determinacy of the future >developments< in the context of objective consequentialism (cf. Harty [2001]; Sinnott-Armstrong [2022] or, once again, the entire actualism-possibilism debate in which the possibility of future decision is explicitly assumed). Prior himself put forward objective probabilities as a means to evaluate these uncertain futures, an approach that we will examine in more detail later and which is still defended today<sup>[127]</sup> and also attacked (cf. Wroński [2020]).

My point is this: This historical trajectory isn't just a theoretical quirk but rather remains a tangible methodological gap. The significance of the under-determination of the future has still *not* seeped into the objective-consequentialist mainstream. However, when objective act-consequentialism is summoned to the complex arena of multi-agent interactions, its traditional focus on individual decision situations with clearly defined consequences for each and every option seems short-sighted. In these collective decision situations, it's evident that the context in which one agent decides can be profoundly influenced, or even sculpted, by the decisions of others.

---

<sup>[127]</sup>For instance, Jackson considered a probabilistic solution (deliberately leaving open the specifics of the relevant account of probabilities) to the actualism-possibilism challenge (cf. Jackson [2014]).

Objective consequentialism must create room for uncertainty. To say it with John Donne (cf. Donne [1923]): *No agent is an island*. And in view of METHODOLOGICAL INDETERMINISM, all agents are to be considered free in their choice. A large number of independent agents with (assumed) freedom of choice and simultaneous mutual dependence with regard to what can be achieved implies a dynamic that cannot be dealt with in any static framework. Thus, uncertainty must have a place in any serious objective consequentialist framework to accommodate multi-agent scenarios, i.e., collective decision situations.

So, the idea underlying APPROACH, i.e., accounting for this kind of non-epistemic but purely theoretical-methodological uncertainty by incorporating decision situations as intermediate outcomes in MOAC's framework, is by no means an esoteric or exotic approach. Instead, it corresponds to the already lived consequentialist practice, well-anchored in preliminary work on decision theory. In order to get a better understanding and feeling for the implications of APPROACH, it is worth developing next a new way of representing collective decision situations under the perspective of the APPROACH.

## 7.2 Towards a Unified Representation: The Generalized Extensive Form

To systematically examine how MOAC might exploit the freedom revealed by APPROACH, a more informative representation of COORDINATION CASES is helpful. This representation should explicitly carve out a space for assessing the novel intermediate outcomes proposed by APPROACH, aiming to offer a cohesive visualization of decision situations. Furthermore, it would be welcome if SEQUENTIAL CASES could also be represented immediately in a corresponding formal structure. Then, we would have a structure that might allow us to find a solution to both the REAL CHALLENGE and the CHALLENGE for COORDINATION CASES as well as for SEQUENTIAL CASES in one go.

While the normal form representation simplifies matters for elementary cases, its general applicability is limited. Given the APPROACH, initial actions are akin to selecting rows or columns in a two-agent scenario, yielding intermediate outcomes. Subsequent actions then pinpoint final outcomes. This ›in-zooming‹ view of outcomes is insightful for simple cases, but visualizing more than three agents becomes convoluted. For instance, visualizing a three-agent scenario is analogous to the subsequent dissection of a cube, and the complexity grows with additional agents, each introducing an additional dimension. Although the matrix-like normal form offers computational advantages – similar to adjacency matrices as representations of graphs – it quickly lacks intuitive appeal and becomes hard to grasp for humans.

Historically, the challenge with the extensive form for COORDINATION CASES has been its ability to be depicted in *one singular* graph. The inherent

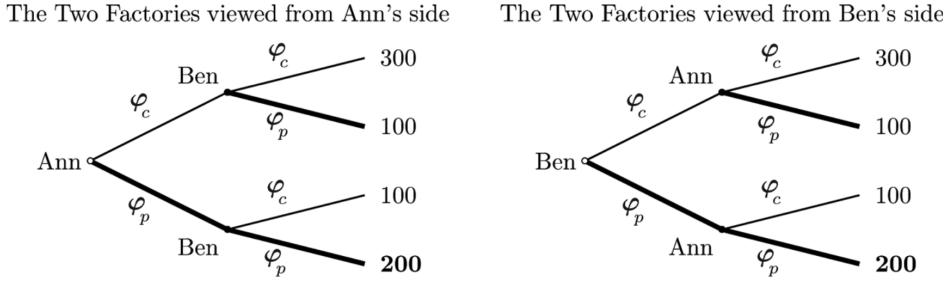


FIG. 2.—The Two Factories as extensive form game

**Figure 7.4:** Pinkert’s *non-unified* representation of his Two FACTORIES case in terms of *two possible* extensive forms that he apparently considers to be one »extensive form game«, cf. Pinkert [2015] p. 975. Note that we would get several more such trees when we allow synchronous actions and want to consider all possibilities.

indeterminacy of the sequence of actions makes the representation inherently ambiguous. To maintain generality, one might contemplate crafting decision trees for *all* possible action sequences, leading to a »forest« of extensive forms (see Figure 7.4). However, with APPROACH offering all necessary intermediate results, we should now be able to disambiguate these possible unfolding in one unified graph. This section, thus, aspires to propose a *generalized extensive form* to navigate the intricacies of the MOAC framework with precision.

In achieving this, retaining simplicity and clarity in the extensive form without omitting MOAC-relevant information is vital. For comparison, consider how the normal form streamlines COORDINATION CASES. Each cell in this format embodies multiple potential outcomes, represented singularly if the outcomes share values and result from combinations of the same actions. For a clearer perspective, consider the Two FACTORIES example: Compare the outcomes of sequences where Ann pollutes before Ben and vice versa. Though technically distinct (there are facts that are true relative to the outcome of the first sequence and false relative to the second, for instance, that Ann acted first), their representation is condensed into a single cell due

to their identical nature in the order-invariant Two Factories scenario.

This suggests the following<sup>128</sup> definition:

**Definition 7.1 (Generalized Extensive Form)**

Let  $D$  be a maximal and order-invariant COORDINATION CASE with  $\mathbf{D} := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi \rightarrow \mathcal{O} \rangle$  (where  $C$  is the actual context of  $D$ ) and  $|\mathcal{A}| = n$ .

The generalized extensive form  $\mathcal{G}(D)$  of  $D$  is a directed acyclic graph, i.e., a tuple  $\mathcal{G}(D) := \langle S_\emptyset \in \mathcal{S}, \mathcal{S}, \mathcal{E} \rangle$  with set of states  $\mathcal{S}$  and a set of transitions  $\mathcal{E}$ .

These sets are defined as follows:

(i) The set of states  $\mathcal{S}$  includes a state for every outcome, be it final or intermediate, from a consequentialist perspective. Given an arbitrary set of representatives  $R \in \mathcal{R}$ , we first define the set of final outcome states as

$$\mathcal{S}_n := \{ S_\gamma \mid \gamma \in R \}.$$

Next, we define recursively

$$\mathcal{S}_{i-1} = \bigcup_{S_\gamma \in \mathcal{S}_i} \{ S_{\gamma \oplus \phi} \mid \phi \in \gamma \}$$

down to  $\mathcal{S}_0 = \{ S_\emptyset \}$ , the start state singleton. Finally, we define the set of states as the union of these sets, i.e.:

$$\mathcal{S} := \bigcup_{i=0}^n \mathcal{S}_i.$$

(ii) The set of transitions  $\mathcal{E}$  is a set of (directed) edges that, in their completeness, represent all sequences of actions in which  $D$  could be resolved. For  $0 \leq i < j \leq n$ , we define

$$\mathcal{E}_{i,j} := \{ \langle S_\gamma, S_{\gamma'} \rangle \mid S_\gamma \in \mathcal{S}_i, S_{\gamma'} \in \mathcal{S}_j, \gamma \sqsubset \gamma' \}.$$

---

<sup>128</sup>This definition uses the notion of a proper part of a combination, i.e., the relation  $\sqsubset$  as defined in Section 5.4.1 on page 235. Further, it makes use of the formerly introduced notions related to the equivalence relation  $\sim^*$  to perform a primitive version of state lumping (Kemeny and Snell 1960), cf. Section 5.2.1.2 on page 216.

*Each of these sets of edges represents all the transitions from all states  $S_\gamma \in S_i$  to the corresponding states  $S_{\gamma'} \in S_j$ . Thus, by construction, each edge represents the combination of action  $\gamma''$  with  $\gamma' = \gamma \oplus \gamma''$  by the corresponding agents (where  $|\gamma''|$  might be one). Finally, we define the set of transitions as the union of all these sets, i.e.,*

$$\mathcal{E} := \bigcup_{i=0}^{n-1} \bigcup_{j=i}^n \mathcal{E}_{i,j}.$$

General extensive forms (shorter, GEFs) can be interpreted as follows. Let  $D$  be a collective decision situation and let  $\mathcal{G}_D = \langle \mathcal{S}, \mathcal{E} \rangle$  be its general extensive form. We can make the following observations.

$S_\emptyset$  represents the starting state of  $D$  and every path from  $S_\emptyset$  to some  $S_\gamma \in S_n$  represents at least one proper combination. In fact, each such path represents all proper combinations  $\gamma' \in [\gamma] \subseteq \Psi$ . All other states in between, i.e., all states  $S_{\gamma''} \in S_i$  (for  $0 < i < n$ ) represent the starting point of reduced decision situations after  $\gamma''$ , which is, by construction, guaranteed to be a proper part of a proper combination, has been performed. (Note that also for each such proper part, there is exactly one such partial path, even though many such paths might be lumped together again as above.)

Accordingly, every edge in  $\mathcal{E}$  corresponds to one or more actions that can happen during  $D$ 's unfolding. While an edge between states  $S_\gamma \in S_i$  and  $S_{\gamma'} \in S_{i+1}$  represents a single action, an edge between two states  $S_\gamma \in S_i$  and  $S_{\gamma'} \in S_j$  with  $j - i = k$  with  $k > 1$  represents the synchronous performance of  $k$  actions.

Finally, every state  $S_\gamma \in \mathcal{S}$  can be seen as the ›defining anchor‹ of a sub-decision situation within  $D$  with a corresponding general extensive form that is a subgraph of the original one. Let  $\gamma \in \Psi_D$  be a proper combination. Then,

for every proper part  $\Upsilon' \sqsubset \Upsilon$  of that combination,  $D$  can be reduced to  $D_{\downarrow \Upsilon'}$ .

Then we can infer  $\mathcal{G}_{\Upsilon'} = \langle \mathcal{S}_{\Upsilon'}, \mathcal{E}_{\Upsilon'} \rangle$  as

$$\mathcal{S}_{\Upsilon'} = \{ S_{\Upsilon''} \mid \Upsilon' \sqsubseteq \Upsilon'' \}$$

and

$$\mathcal{E}_{\Upsilon'} = \{ \langle S_{\Upsilon''}, S_{\Upsilon''' \rangle} \in \mathcal{E} \mid S_{\Upsilon''}, S_{\Upsilon''' \in \mathcal{S}_{\Upsilon'}} \}.$$

In other words, to arrive at  $\mathcal{G}_{D_{\downarrow \Upsilon'}}$ , we throw all parts of  $\mathcal{G}_D$  away that are not consistent with  $\Upsilon'$  has already been performed. We thus have a one-to-one correspondence between intermediate states and the individual decision situations that serve as intermediate outcomes according to APPROACH.

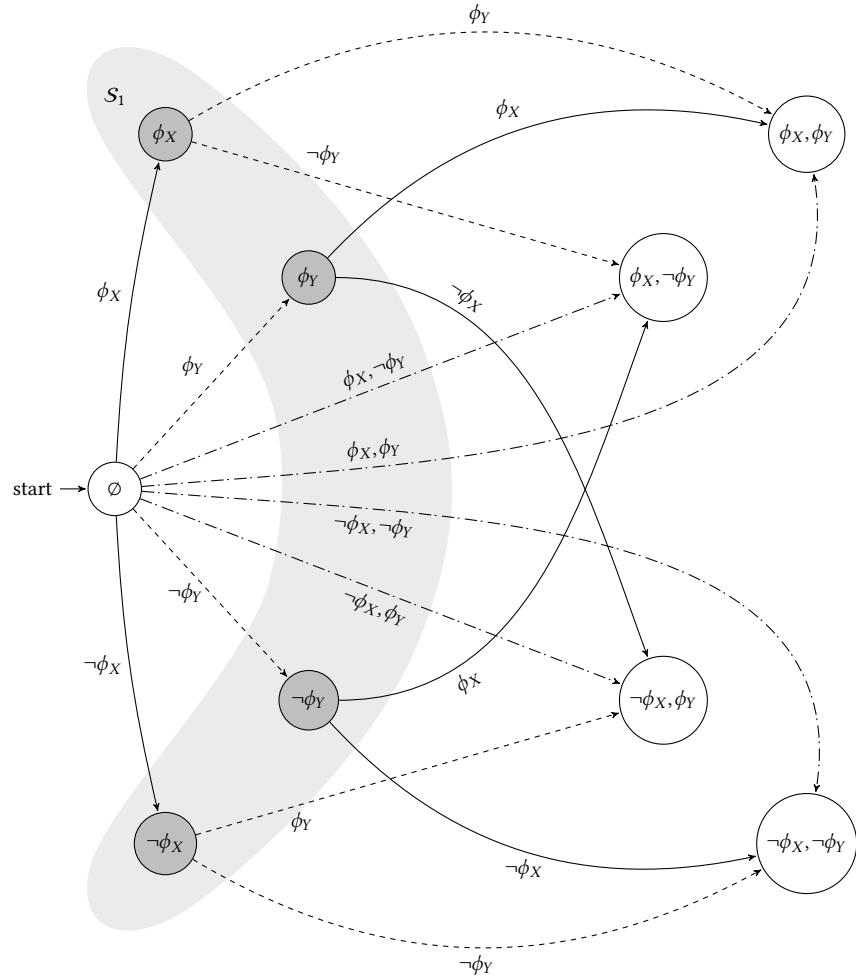
It is useful to have a shorthand for the set of all these intermediate outcomes. First, given a GEF  $\mathcal{G} = \langle \mathcal{S}, \mathcal{E} \rangle$ , we note that the *set of intermediate states*  $\mathcal{S}^{\text{inter}}$  is given, by construction, as:

$$\mathcal{S}^{\text{inter}} = \bigcup_{i=1}^{n-1} \mathcal{S}_i.$$

In other words, given a GEF, we throw away the start state  $S_\emptyset$  and all final outcomes, i.e., those states without outgoing edges. Next, we define the *set of intermediate outcomes* as

$$\mathcal{O}^{\text{inter}} = \{ D_{\downarrow \Upsilon} \mid S_\Upsilon \in \mathcal{S}^{\text{inter}} \}.$$

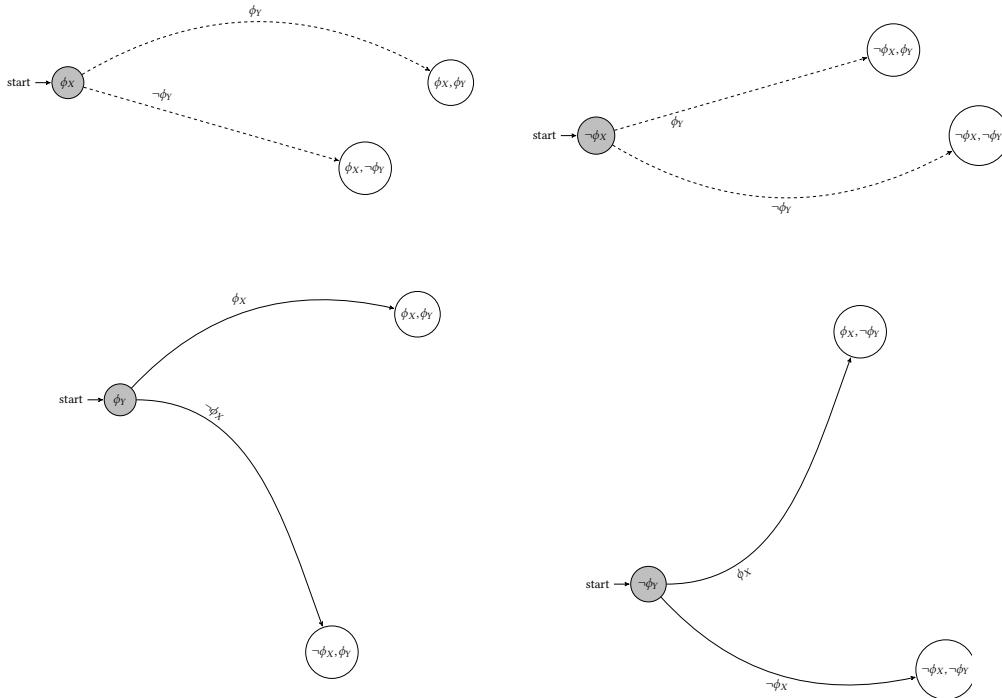
GEFs can get quite big and cluttered quickly even for small COORDINATION CASES. Figure 7.5 shows an example of a GEF for a minimal case, i.e., a maximal and order-invariant COORDINATION CASE with two agents and two options each, i.e., for a COORDINATION CASE we would traditionally present in a normal form like this:



**Figure 7.5:** Generalized extensive form for a maximal, order-invariant COORDINATION CASE with two agents ( $X$  and  $Y$ ) each having two options ( $\phi_z$  and  $\neg\phi_z$  for  $z \in \{X, Y\}$ ). Solid arrows denote actions by  $X$ ; dashed arrows represent actions by  $Y$ ; and dash-dotted arrows indicate synchronous actions by both. States are denoted by their index, e.g., state  $S_\gamma$  is labeled  $\gg\gamma\ll$ . White-background states (excluding the starting state) show traditional results such as  $\text{Out}(\langle\phi_X, \phi_Y\rangle)$ , lumped following the above-defined procedure (note that otherwise, we would have eight such white states). Dark gray states are intermediate, resulting from either  $X$ 's row choice or  $Y$ 's column choice, collectively forming the set  $S_1$  (shown within the boomerang-shaped light-gray zone). These states are starting points for four sub-graphs (see Figure 7.6), resulting from actions corresponding to the edge leading to these states.

		$Y$	
		$\phi_Y$	$\neg\phi_Y$
		$\phi_X$	$\text{Out}(\phi_X, \phi_Y)$
$X$		$\neg\phi_X$	$\text{Out}(\phi_X, \neg\phi_Y)$
		$\phi_Y$	$\text{Out}(\neg\phi_X, \phi_Y)$
		$\neg\phi_Y$	$\text{Out}(\neg\phi_X, \neg\phi_Y)$

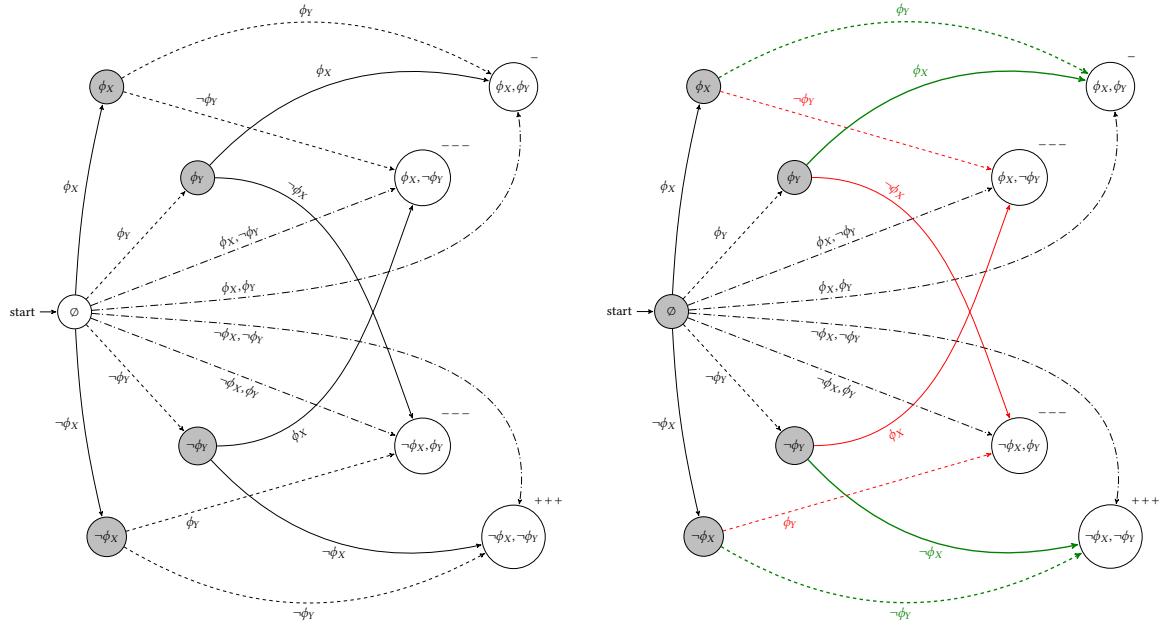
One can see that the GEF of this case allows us to distinguish more



**Figure 7.6:** The four sub-graphs, each corresponding to  $D_{\downarrow\phi_X}$ ,  $D_{\downarrow\neg\phi_X}$ ,  $D_{\downarrow\phi_Y}$ , and  $D_{\downarrow\neg\phi_Y}$ , respectively (from left to right and top to bottom), and thus to one of the ›new outcomes‹ according to APPROACH.

clearly the relationships between the different combinations of actions and their occurrence (successive or simultaneous). Also, we can see at a glance the intermediate outcomes (cf. Figure 7.6), which are the same kind of thing as the GEF of the collective decision situation, namely a (sub-)graph.

At this point, thus, we can begin to *reason morally about paths* from the start to the final outcomes when thinking about where morality *should* (in a sense still to be defined in more depth) ›guide‹ the agents within a collective decision situation. Let us consider the following generic Two FACTORIES-like case that we get from the above case by adding a fitting value profile:



**Figure 7.7:** The GEF of Two Factories-like cases with value annotations (added top right of the states representing final outcomes, i.e., white states). On the right with additional moral assessment annotation (based on MOCoR): green for edges representing right actions and red for edges representing wrong actions. For actions corresponding to gray actions, MOCoR is not yielding assessments, i.e., only future-involving conditional assessments are available. (In addition, we can derive individual backward-looking assessments in Pinkert's style, cf. Section 3.5.2.3, page 139.)

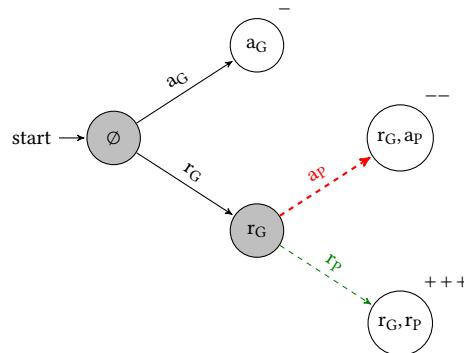
		Y	
		$\phi_Y$	$\neg\phi_Y$
$\phi_X$	$\phi_X$	—	---
	$\neg\phi_X$	---	+++

We can add these valuations directly to the GEF (cf. Figure 7.7, left side, and for JOB MARKET in Figure 7.8<sup>129</sup>). In the next step, we can then try to apply MOCoR. This gives us a visualization of both the CHALLENGE and the REAL CHALLENGE (cf. Figure 7.7, right side). We recognize the CHALLENGE in the fact that there are green edges towards the  $\phi_X, \phi_Y$  state. As soon as we end up in that state and we start the counterfactual reasoning

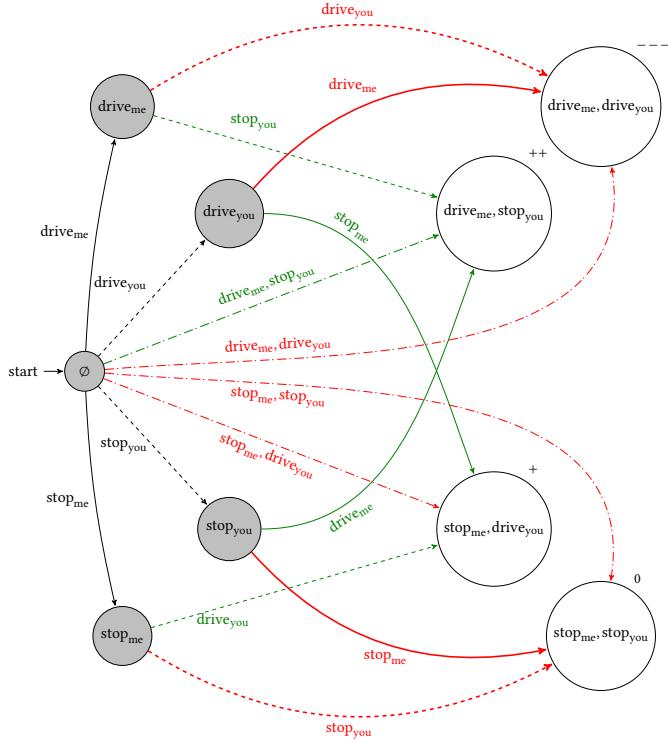
<sup>129</sup>For SEQUENTIAL CASES the concept of GEFs doesn't change much but we can easily see that we now have a truly unified representation for both kinds of cases.

typical for the CSM, we find – correctly – that if any of the agents had acted differently, they would only have made things worse, i.e., acted wrongly and, hence, followed a red edge. The CHALLENGE, as analyzed in the last chapter, thus has its origin in that the champions of MOAC, to get assessments at all, move right to the final state and then, when arguing for the rightness of both agents' actions that lead there, only moves back to the  $\phi_X$  or  $\phi_Y$  state (depending on whether they are pondering on  $\phi_Y$  or  $\phi_X$ ).

The REAL CHALLENGE is visualized insofar that we now recognize that a violation of PMH would happen only when we found a green path from the starting state to the  $\phi_X, \phi_Y$  state. But no such path exists, simply because *no green edge leaves the start state at all*. The REAL CHALLENGE is precisely this: It is not the case that a green path leads to a suboptimal state; instead, there is no completely green path leading anywhere; especially there is no green path to the optimal state. According to MOAC, there simply is *no right first step* in any direction. Thus, there is no possibility of following the commandments and recommendations of MOAC (as there are none). All there is... are deontic gaps.



**Figure 7.8:** The (unspectacular) GEF of JOB MARKET with annotations where »*a*« stands for »accepting the job offer«, »*r*« stands for »rejecting the job offer«, »*G*« stands for »George«, and »*P*« stands for »Paul«. Extensive forms just remain extensive forms. Yet, we see that also, in this case, we only get assessments for Paul's (hypothetical) decision situation – and that it seems more than plausible to say that George's option of rejecting the job has an individual decision situation as a consequence, namely Paul's. Arguably, the same line of reasoning that resolves COORDINATION CASES should resolve SEQUENTIAL CASES automatically as COORDINATION CASES are, in light of the APPROACH, just *superpositions* of several SEQUENTIAL CASES and solving a COORDINATION CASES comes down to solving all its constituting SEQUENTIAL CASES.



**Figure 7.9:** The GEF of INTERSECTION with annotations according to Jackson’s reasoning. We can observe how the combination of me driving and you stopping is right according to his reasoning in the initial state. However, once you drive, it is right for me to stop (*because* you drive). This is another decision situation and, thus, another option. Hence, it is highly misleading the least to say that the »right combination« of me driving and you stopping *contains* a wrong action, namely me driving. The combination is right (in Jackson’s terms) relative to the initial state; me driving is wrong relative to the state labeled »drive<sub>you</sub>«.

Before I move on and discuss how consequentialists might close these gaps, we can reconnect to the end of the first part: It is now easy to explain what was wrong with Jackson's reasoning in Section 4.4.3. The *frame of reference* just wasn't right. Let me explain.

There are good reasons to believe that the combination of the action of you to stop and me to drive is, in a sense, right. While this is to be negotiated in detail in the course of the rest of this book (because we do not yet know how to assess the actions starting from the initial state), we now understand enough to get to the heart of the flaw in Jackson's reasoning. As soon as you drive, 'we' are in a different state and, thus, in a different decision situation. From this new context, there are only options for me to choose from as it's *my*

remaining individual decision situation (you and your action are just part of the context now). And, of course, it is then right for me to stop – just *because* you are driving and, therefore, I am in the said individual decision situation as a result of the reduction of the original collective decision situation of us two.

However, and this is the crucial point, by this, *nothing at all* is said about what individual options are right in the initial state. The »because« makes an essential difference with respect to the frame of reference: I am in another decision situation than we were and whatever is right or wrong for me now, says nothing about what was right or wrong to do in the collective decision situation, from which we come (cf. Figure 7.9). To assume something else corresponds *exactly* to the collective original sin of consequentialism, i.e., to CSM, which wanted to close the deontic gaps in that way and, *by doing so*, ran so into the CHALLENGE (cf. Chapter 6). Jackson's observation is thus a bug, not a feature, in the standard framework of the consequentialists that is too shallow and too poor in structure. Or it was – until today.

### 7.3 Filling Gaps With Multi-Agent Amendments

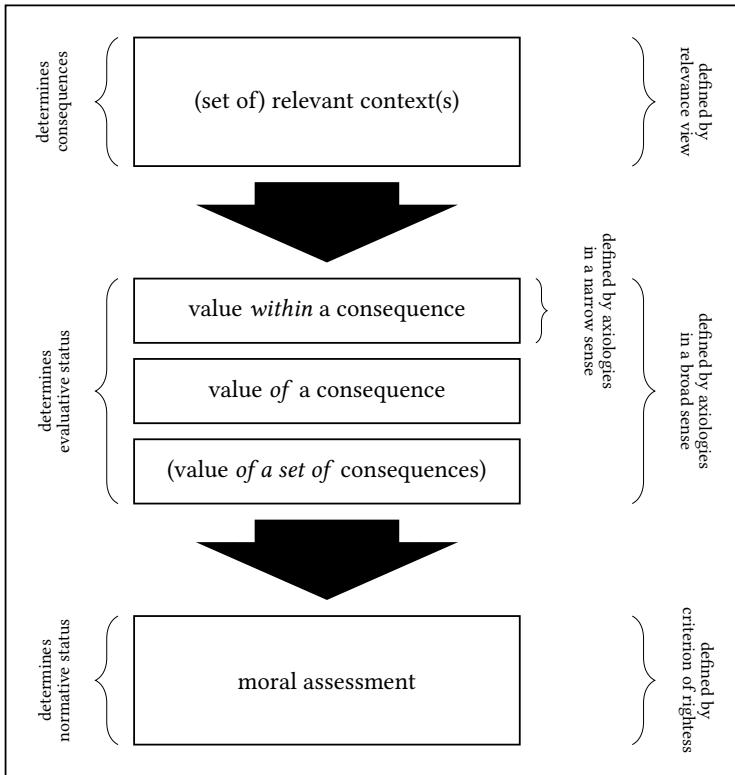
Given the APPROACH, the solution of the REAL CHALLENGE seems within reach. The defensibility of COMPOSITIONALISM has been achieved, and we have identified an individual decision situation for each agent in each collective decision situation, including proper consequences for each of the agents' options. Camp MOAC thus merely needs a matching puzzle piece to close the gaps, i.e., a way to assess these options morally to fit MOAC's ambitions. As mentioned before, I call these theoretical extensions to the consequentialist toolbox *multi-agent amendments*.

As it turns out, there are many *prima facie* promising candidates. Unsurprisingly, they all have in common that they ultimately make use of the already given evaluations of the final outcomes. This makes them consequentialist amendments.

Two classes of amendments can be distinguished. One is *aggregative*. They are extensions of the axiological toolbox in a broad sense (cf. Section 2.3.3), adding a third dimension of aggregation to the two traditional aggregation dimensions over time and over moral patients. Thus, these methods are ›interpossible‹, aggregating over how things might possibly unfold. Some of these methods are probabilistic and involve objective probabilities – or, at any rate, can be extended accordingly. Aggregative amendments allow us to rank the newly discovered outcomes, i.e., the remaining decision situations of the other agents, in such a way that MOCoR can be applied without a need for any modification.

The other class of amendments is *non-aggregative*. They get by without evaluating the corresponding intermediate outcomes and still arrive at a *ranking of options based on their consequences moral quality*. Thus, the resulting overall theories remain, in a sense, MOAC theories, even though they might make a modification of MOCoR necessary.

In the following, I present a selection of amendments that I consider to be particularly promising and theoretically well-anchored. Most are based on methods for decision-making under uncertainty or under risk (cf. Section 7.1.2). The question of how Camp MOAC should ultimately decide between these amendments is then the subject of the following chapter.



**Figure 7.10:** How the modules of a consequentialist theory interlink to arrive at moral assessments. Aggregative Amendments simply give MOAC theories a way to evaluate sets of consequences, as it is common for subjective varieties of consequentialism.

### 7.3.1 Aggregative Approaches

Aggregative amendments extend the valuation function  $\text{Val}$ . While  $\text{Val}$  has so far been defined over final outcomes (and we have syntactically defined, as shorthand,<sup>130</sup> also a direct application to options), to benefit MOCoR from the new intermediate outcomes that the APPROACH has brought to the work-bench of consequentialism,  $\text{Val}$  must be able to value these very outcomes. That is, it must be able to evaluate both individual *and* collective decision situations (because in cases with more than two agents, the consequences of the first-acting agent's action remain a collective decision situation, albeit

<sup>130</sup>Recall that, given an individual decision situation  $D$  with actual context  $C$ ,  $\text{Val}(\phi)$  for an  $\phi \in \Phi_D$  was simply defined as shorthand for  $\text{Val}(\text{Out}_C(\phi))$ , cf. 51.

only one of the two remaining agents). This approach to closing the gaps fits nicely into the picture of the action-consequentialist evaluation pipeline painted in Section 2.3. Ultimately, aggregative amendments merely insert an evaluation for sets of possible outcomes, just as it is common for subjective varieties of consequentialism (cf. Figure 7.10).

Before we get into the candidates for aggregative amendments, it is worth recalling some notational trivia. First, recall that, thanks to APPROACH, we can decompose any arbitrary collective decision situation  $D$  with  $D := \langle \mathcal{A}, \Gamma, \text{Out}_C : \Psi_\Gamma \rightarrow \mathcal{O} \rangle$  and actual context  $C$  into  $n := |\mathcal{A}|$  individual decision situations  $D_{A_1}, \dots, D_{A_n}$  with the structure

$$D_{A_i} := \langle A_i, \Phi_{A_i}, \text{Out}_C : \Phi_{A_i} \rightarrow \mathcal{O}_D^{A_i} \rangle$$

where

$$\mathcal{O}_D^{A_i} := \{ D_{\downarrow \phi} \mid \phi \in \Phi_{A_i} \}.$$

Let us call the set

$$\mathbb{I}_D := \{ D_{A_i} \mid \mathcal{A}_D \}$$

the *decomposition of D*.

Further, recall that we defined  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  where

$$\mathcal{W} := \bigcup_{D \in \mathbb{I}} \mathcal{O}_D$$

(For simplicity, I assume  $\mathcal{V} = \mathbb{R}$ , for the rest of this project, although less restrictive constraints would suffice in principle.) Finally, be reminded that we introduced the short hands  $\text{Val}(\phi)$  and  $\text{Val}(\Upsilon)$  for  $\text{Val}(\text{Out}(\phi))$  and  $\text{Val}(\text{Out}(\Upsilon))$ .

With these in mind, it is time to introduce a first amendment.

### 7.3.1.1 SUMMATION

The two methods of aggregation across moral patients and points in time that are typically associated with MOAC are both simple summations. It is, therefore, worth starting with the following simple and straightforward candidate amendment:

**Definition 7.2 (SUMMATION)** *Let  $D$  be a individual decision situation with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  and actual context  $C$ . Further, let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The value of a individual decision situation  $D$  is*

$$\text{Val}^\Sigma(D) := \sum_{\phi \in \Phi} \text{Val}(\phi).$$

*Let  $D$  be a collective decision situation with actual context  $C$  and decomposition  $I_D := \{ D_{A_i} \mid A_i \in \mathcal{A}_D \}$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.*

*The value of a collective decision situations to  $D$  is*

$$\text{Val}^\Sigma(D) := \sum_{D_{A_i} \in I_D} \text{Val}^\Sigma(D_{A_i}).$$

So, we first lift the valuation function  $\text{Val}$  from the level of outcomes to the level of individual decision situations by defining the value of an individual decision situation as the sum of the values of its possible outcomes, i.e., by simply summing up all possible outcomes of the situation. Then we recursively define the value of a collective decision situation as the sum of the values of its individual decompositions. This recursive definition is guaranteed to be well-defined for cases with a finite number of agents (and decisions) because their value is, again, the sum of the values of the possible outcomes of those individual decision situations. Since these are, in the worst case, even collective decision situations with one less agent, we reduce the remaining

decision situations further and further until we arrive at the valuation of final outcomes and thus reach a termination.

We remember our running example **TWO FACTORIES**. Recall

		Ben	
		pollute	produce cleanly
		pollute	second-worst
Ann	pollute	second-worst	worst
	produce cleanly	worst	best

If we assume a few concrete values, we can now directly determine the valuation of options of Ann according to  $\text{Val}^\Sigma$ :

		Ben		$\text{Val}^\Sigma(\cdot)$
		pollute	produce cleanly	
Ann	pollute	-1000	-2000	-3000
	produce cleanly	-2000	+1000	-1000

So we see that according to MOAC along with the SUMMATION amendments, it would be unconditionally right for Ann in **TWO FACTORIES** to produce cleanly because -1000 is larger than -3000. We can do exactly the same for Ben, of course:

		Ben		$\text{Val}^{\Sigma}(\cdot)$
		pollute	produce cleanly	
Ann	pollute	-1000	-2000	-3000
	produce cleanly	-2000	+1000	-1000
$\text{Val}^{\Sigma}(\cdot)$		-3000	-1000	

We see that MOAC together with SUMMATION → leads to the best result in this case.

This is not a contingent finding for the specific values chosen. We can represent TWO FACTORIES more generally as

		Ben		$\text{Val}^{\Sigma}(\cdot)$
		pollute	produce cleanly	
Ann	pollute	$v_{pp}$	$v_{pc}$	$v_{pp} + v_{pc}$
	produce cleanly	$v_{cp}$	$v_{cc}$	$v_{cp} + v_{cc}$
$\text{Val}^{\Sigma}(\cdot)$		$v_{pp} + v_{cp}$	$v_{pc} + v_{cc}$	

where  $v_{cp} = v_{pc} < v_{pp} < v_{cc}$ . Even more generally we have

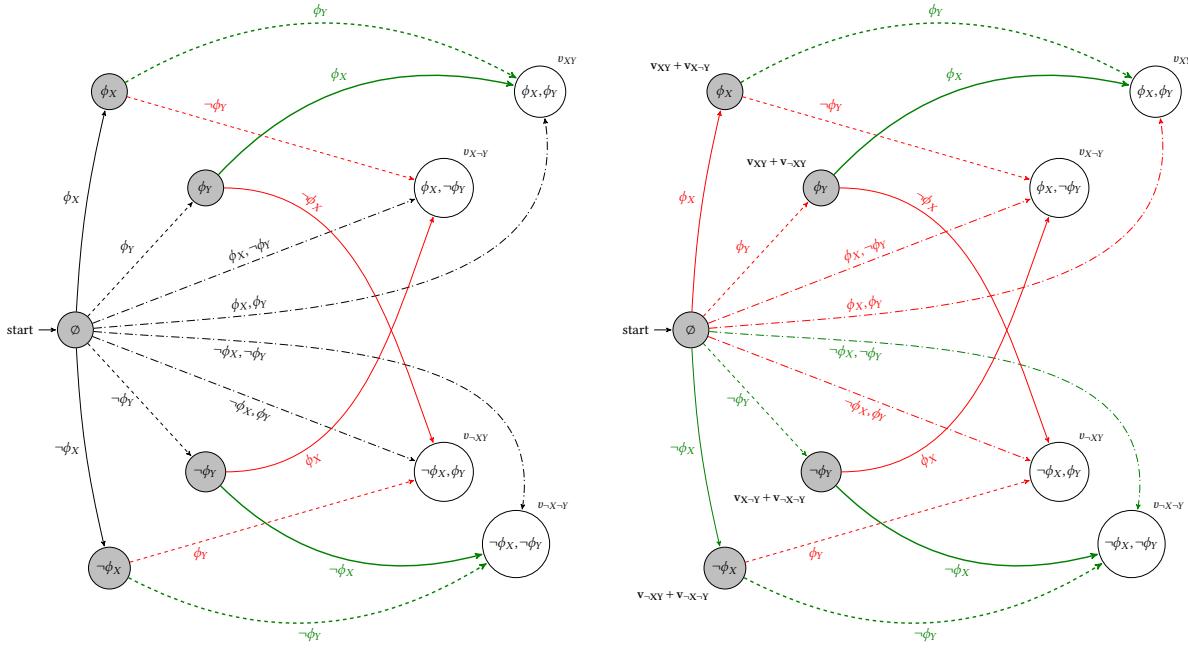
	$Y$		$\text{Val}^\Sigma(\cdot)$
	$\phi_Y$	$\neg\phi_Y$	
$\times$	$v_{XY}$	$v_{X-Y}$	$v_{XY} + v_{X-Y}$
	$v_{\neg XY}$	$v_{\neg X-Y}$	$v_{\neg XY} + v_{\neg X-Y}$
$\text{Val}^\Sigma(\cdot)$	$v_{XY} + v_{\neg XY}$	$v_{X-Y} + v_{\neg X-Y}$	

where  $v_{X-Y} = v_{\neg XY} < v_{XY} < v_{\neg X-Y}$ . Thus, we know that  $v_{XY} < v_{\neg X-Y}$ . Since  $v_{X-Y} = v_{\neg XY}$ , we can rewrite that inequality to both  $v_{XY} + v_{\neg XY} < v_{X-Y} + v_{\neg X-Y}$  as well as to  $v_{XY} + v_{X-Y} < v_{\neg XY} + v_{\neg X-Y}$ . Therefore, it is true that  $\text{Val}(\phi_z) < \text{Val}(\neg\phi_z)$  (for  $z \in \{X, Y\}$ ) independent from the concrete choice of values.

So we can say that MOAC with SUMMATION yields for TROUBLE-MAKERS like WHIFF AND POOF and Two FACTORIES assessments quite in the spirit of PMH. Figure 7.11 shows how the gaps have been closed in Two FACTORIES's annotated GEF.

As a side effect, we can now give a plausible answer to the question of the borderline case of simultaneous action: we only have to look at the actions of the individual agents and their assessments in the initial state. From this, the indirect assessment of the action combinations results arguably in a way that respects METHODOLOGICAL INDIVIDUALISM: Only the combination of actions which consists exclusively of right actions is right in the derived sense.

I will not carry out analogous considerations in this generality and level of detail for the following amendments. Instead, I will restrict myself to evaluating one or two example cases. It stands to reason, however, that we can



**Figure 7.11:** The GEF of Two Factories-like cases with value annotations and moral assessment annotation (based on MOCoR and under the assumption that  $v_{X-Y} = v_{-XY} < v_{XY} < v_{-X-Y}$ ): green for edges representing right actions and red for edges representing wrong actions. For actions corresponding to gray actions, MOCoR is not delivering assessments, i.e., only future-involving conditional assessments are available. Left for MOCoR only; right for MOCoR with SUMMATION as the amendment of choice (the induced value at the intermediate states are set in bold). Note how the only fully green paths lead to the global optimum.

learn something from this insight about how camp MOAC should choose among the various amendment candidates, namely by looking at the results of this choice relative to classes of collective decision situations.

### 7.3.1.2 MAXIMIZATION

Next, we consider another type of aggregation as an amendment, which, in a sense, explicitly honors the »M« in MOAC. The basic idea is to approach the task of closing the gap by thinking from the end. Instead of starting the considerations from the initial state and where the first actions might lead, one starts with the value of the final outcomes. One then assesses all combinations' actions according to the best final outcome they can lead to.

To get to the heart of this idea, we use an auxiliary definition: Let  $D$  be a collective decision situation with domain  $\Psi$  and actual context  $C$ . Further, let

$\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. For an arbitrary  $\phi \in \Phi_A \in \Gamma_D$ , we define the set of combinations that involve  $\phi$  as:

$$\Psi_\phi := \{ \Upsilon \in \Psi \mid \phi \in \Upsilon \}$$

Based on this set, we can now ›push down‹ the valuation function from the level of combinations to individual actions.

**Definition 7.3 (MAXIMIZATION proto)** *Let  $D$  be a collective decision situation with actual context  $C$  and decomposition  $\mathbb{I}_D := \{ \mathsf{D}_{A_i} \mid A \in D \}$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.*

The value of a individual option  $\phi \Phi_A$  of some  $A$  in  $D$  is

$$\text{Val}^{\max}(\phi) := \max_{\Upsilon \in \Psi_\phi} \text{Val}(\Upsilon).$$

This formulation of the idea, noteworthy, comes without recourse to the APPROACH. Thus, in this form, MAXIMIZATION seems rather ad-hoc from a consequentialist point of view, because, suddenly, the moral assessment of an action is no longer a direct function of the moral quality of its consequences (but only of a specific *possible* consequence in the light of the actions of other agents' actions).

With APPROACH backing us up, however, we can get around this issue. Instead of modifying the valuation function from combinations to options without certain consequences, we can extend it to decision situations, much as we did with SUMMATION:

**Definition 7.4 (MAXIMIZATION)** *Let  $D$  be a individual decision situation with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  with actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The value of a individual decision situation  $D$  is*

$$\text{Val}^{\max}(D) := \max_{\phi \in \Phi} \text{Val}(\phi).$$

Let  $D$  be a collective decision situation with actual context  $C$  and decomposition  $I_D := \{ D_{A_i} \mid A_i \in \mathcal{A}_D \}$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.

The value of a collective decision situations to  $D$  is

$$\text{Val}^{\max}(D) := \max_{D_{A_i} \in I_D} \text{Val}^{\max}(D_{A_i}).$$

The resulting ›gap filling‹ based on these two definitions of MAXIMIZATION is indeed extensionally equivalent.<sup>131</sup> However, we now no longer need to adapt MOCOR because we simply assess the direct consequences of the respective options established thanks to APPROACH.

For illustration, we can apply this amendment to TROUBLEMAKERS. This time, we replace the generic notation with a notation of »+« and »-« signs, which makes the order of the values of the final outcomes immediately evident.

	Y		$\text{Val}^{\max}(\cdot)$
	$\phi_Y$	$\neg\phi_Y$	
$\phi_X$	-	---	-
$\neg\phi_X$	---	+++	+++
$\text{Val}^{\max}(\cdot)$	-	+++	

<sup>131</sup>I leave the proof to the interested reader, but the intuition should be clear: The value of an individual action is equal to the value of its consequence (this corresponds to the shorthand notation we have agreed upon). According to APPROACH, this consequence of an individual agent's action is the reduced collective decision situation relative to this very action. If that reduced situation is an individual decision situation, then, according to the first part of Definition 7.4, the value of this situation is the value of the best final outcome reachable from it. If that reduced situation is collective, then, according to the second part of Definition 7.4, the value of this situation is the value of the best decision situation reachable from it. Ultimately, we will end up in case 1 because we consider only situations with a finite number of agents. Thus, the end result is that the value of each option in the original collective situation corresponds to the value of the best final outcomes reachable by any combination containing that option – which corresponds exactly to Definition 7.3

We thus see that MOAC together with MAXIMIZATION is again, like it was in combination with SUMMATION, guaranteed to lead to the best possible result for all TROUBLEMAKERS with this valuative profile. As with SUMMATION this is no contingent finding as it is independent of the specific values of the final outcomes, but it is implied by the order of these values.

### 7.3.1.3 EXPECTED UTILITY

The following amendment is motivated by decision-theoretic MSAC and attempts to make the concept of expected value fruitful for objective consequentialists. Thus, this approach can also be understood as the application of Kagan's solution approach to MOAC theories (cf. Section 4.4.1). The most important *theoretical* drawback of this approach is that objective probabilities are needed both in terms of what actions each agent will perform and in terms of what pure order the agents will act. On the one hand, it is utterly unclear where to get suitable probability distributions from, on the other hand, objective probabilities are also metaphysically and ontologically a somewhat questionable kind of thing. However, as we will see, there is also an *extensional* drawback, that is, to all appearances, there are good reasons why MOAC should not adapt EU.

The advantage, however, is that the approach would make it possible to incorporate the actual or expected actions of the other agents in the process of moral assessment of some agent's options, allowing MOAC to make conceptual space for risks. Suppose Ann has already had many experiences with Ben and knows that Ben will pollute. Alternatively, suppose that Ben has a track record of environmentally destructive behavior. Let us assume that all this is evidence of a certain bad disposition of Ben. Beyond the question of

what Ann should do from a subjective point of view (in a moral sense), the question arises as to whether an appropriate extension of MOAC to multi-agent scenarios should really make the consistently right-acting of all agents the basis of its assessments. Because then, acting right would often mean running into really bad and otherwise avoidable moral disasters. In favor of a version of MOAC based on expected values speaks what also speaks for it in subjective settings: A certain balance of risk awareness, nevertheless aiming at some kind of optimum-oriented decision-making.

I am by no means the first to have this idea. In the essay already quoted in the last chapter, A.N. Prior (A. N. Prior and Raphael [1956], pp. 93) writes, basically in response to the there-sketched REAL CHALLENGE:

Taking the non-determinist horn first, perhaps we can say that if determinism is not true, it suffices to speak of a duty to do what will *probably* have the best total consequences of all the actions open to us. We can only take this line, however, if we are prepared to talk about objective probabilities ; that is, if we are prepared to argue that »*p* is probable« need not merely mean »We don't know that *p* will be true, but what evidence we have is more in favour of it than against it«, but may mean something more like »*p* is not yet either going to be the case or not going to be the case, but is more like going to be the case than not«.

I think that the burden of objective probabilities brings too much theoretical ballast for MOAC. While I do not want to exclude that consequentialists find a way to carry this burden and make it fruitful in the practice of normative ethics, I do not make it my project to defend this approach here. So, I merely sketch here what the corresponding amendment would look like and then shelve it.

As already mentioned, for each individual decision situation, we need a probability distribution of which action will be taken. Further, we need for each collective decision situation a probability distribution of which agent will act when.<sup>132</sup> In the following, I use  $\Pr$  for these distributions, each with an index indicating the set over which the distribution runs. We can then define:

**Definition 7.5 (EXPECTED UTILITY)** *Let  $D$  be a individual decision situation with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  with actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The (expected) value of a individual decision situation  $D$  is*

$$\text{Val}^{EV}(D) := \sum_{\phi \in \Phi} \Pr_{\Phi}(\phi) \text{Val}(\phi).$$

*Let  $D$  be a collective decision situation with actual context  $C$  and decomposition  $I_D := \{ D_{A_i} \mid A \in \mathcal{A}_D \}$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The (expected) value of a collective decision situations to  $D$  is*

$$\text{Val}^{EV}(D) := \sum_{D_{A_i} \in I_D} \Pr_{\mathcal{A}}(A \text{ acts next}) \cdot \text{Val}^{EV}(D_{A_i}).$$

*The (expected) value of an option  $\phi$  within a collective decision situations to  $D$  is*

$$\text{Val}^{EV}(\phi) := \text{Val}^{EV}(D_{\downarrow \phi}).$$

Since I don't explore EXPECTED UTILITY further in this book, I don't go into much detail about what this amendment implies with respect to TROU-

---

<sup>132</sup>This second probability distribution could also be of a less metaphysical nature, i.e., we do not necessarily have to assume that it represents some kind of probabilistic fact. Instead, we might suggest that every sequence of actions is equally probable (where we would possibly want to regard simultaneous actions as quasi-impossible at sufficiently precise 'temporal resolution'). But even such an assumption would require justification, which I cannot provide here, at least not against the background of an EXPECTED UTILITY-focused narrative (indeed, I will revisit this rather analytic approach in a different context in the next chapter).

BLEMAKERS. However, it is worth looking briefly at the following instance of TWO FACTORIES:

		Ben		$\text{Val}^{\text{EV}}(\cdot)$
		0.8	0.2	
		pollute	produce cleanly	
Ann	0.8	-1000	-2000	-1200
	0.2	-2000	+1000	-1400
$\text{Val}^{\text{EV}}(\cdot)$		-1200	-1400	

What this instance shows is that the result of *following*  $\text{Val}^{\text{EV}}$  stringently can be suboptimal (because, given the chosen probability distributions, producing in a polluting manner has a higher expected value for both agents than producing clean, so both would act rightly if they produced dirty). So, in essence, filling the deontic gaps using  $\text{Val}^{\text{EV}}$  seems to tend to provide room for the CHALLENGE to re-emerge.

What we also see is that the non-conditional assessments of  $\text{Val}^{\text{EV}}$  amended MOCOR depend crucially on which concrete probability distributions prevail.<sup>133</sup> In other words, the traditional descriptions of collective action situations are strongly incomplete relative to this approach (which is unsurprising since objective probabilities in the context of MOAC are an unusual thing to consider). Moreover, we'd need to overdo our definition of symmetry by incorporating probability distributions alongside the valuative profile. In conclusion, although there might still be potential for EXPECTED UTILITY to be a viable position within camp MOAC, adopting it would introduce substantial theoretical complexities. At the same time, there's no guarantee that

<sup>133</sup>The assessments above would already switch, for instance, for a 75% : 25% distribution for both agents.

this would even help to address the CHALLENGE effectively. Therefore, I set aside EXPECTED UTILITY.

Of course, there are infinite other ways to evaluate the consequences newly gained thanks to APPROACH in an aggregative manner. However, I believe that the here-presented amendments are the most important ones that deserve closer consideration given established methods from decision and game theory and in light of other consequentialist defaults. So, let us turn to non-aggregative amendments next.

### 7.3.2 Non-Aggregative Amendments

Non-aggregative amendments allow us to rank the individual options of agents in collective action situations in terms of consequentialism, that is, based only on the moral quality of immediate consequences that the APPROACH offers, but *without evaluating* these consequences. The advantage is that, strictly speaking, many of these amendments work without the APPROACH and, thus, do not require an extension of axiology in the broad sense. The disadvantage, however, consists in the need to adapt MOCoR for them to get traction. Further, without recourse to the APPROACH, these adaptations quickly seem *ad-hoc* as the link between the assessed option and the specific consequences considered becomes at least blurred.

In the following, I introduce five of many conceivable amendments, which seem to me to be the most obvious and promising ones. They fall into three groups.

### 7.3.2.1 (Non-)Domination

Based on the notion of  $\Psi_\phi$  from above, we can define what it means that one combination dominates the other. For two options  $\phi, \phi' \in \Phi_A \in \Gamma_D$  of the same agent, we write  $\phi \geq \phi'$  for that  $\phi$  weakly dominates  $\phi'$  and write  $\phi > \phi'$  for that  $\phi$  strongly dominates  $\phi'$ . We define:

$$\phi \geq \phi' \text{ if and only if } \forall \Upsilon \in \Psi_\phi, \forall \Upsilon' \in \Psi_{\phi'} : \text{Val}(\Upsilon) \geq \text{Val}(\Upsilon')$$

and, accordingly,

$$\phi > \phi' \text{ if and only if } \forall \Upsilon \in \Psi_\phi, \forall \Upsilon' \in \Psi_{\phi'} : \text{Val}(\Upsilon) > \text{Val}(\Upsilon')$$

This allows us to define two<sup>134</sup> collective criteria of rightness:

**Definition 7.6 (DOMINANCE)** *Let  $D$  be a collective decision situation with actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.*

*The option  $\phi \in \Phi_A$  of some  $A$  in  $D$  is right for  $A \in \mathcal{A}_D$  in  $C$  if and only if  $\phi$  weakly dominates every other option of  $A$  in  $C$ , i.e., if and only if  $\forall \phi' \in \Phi_A : \phi \geq \phi'$ .*

**Definition 7.7 (DOMINANCE-free)** *Let  $D$  be a collective decision situation with actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.*

*The option  $\phi \in \Phi_A$  of some  $A$  in  $D$  is right for  $A \in \mathcal{A}_D$  in  $C$  if and only if  $\phi$  is not strongly dominated by any other option of  $A$  in  $C$ , i.e., if and only if  $\nexists \phi' \in \Phi_A : \phi' > \phi$ .*

---

<sup>134</sup> DOMINANCE-free corresponds to what John F. Harty once called »dominant act utilitarianism«, cf. Harty [2001], chapter 4. To the best of my knowledge, Harty's purely formal work on multi-modal deontic logic, limited to stit semantics, is the closest thing to my project. My thoughts have developed independently of Harty's book, which I came across only later through a direct conversation with him at a Dagstuhl seminar in 2016. I am deeply indebted to Jeff and his work, though.

Both of these criteria offer distinct moral guidance in collective scenarios. Reconsider our standard example with the structure of a TROUBLEMAKER:

		$Y$			
		$\phi_Y$	$\neg\phi_Y$	is dominant	is dominance-free
$X$	$\phi_X$	—	---	false	true
	$\neg\phi_X$	---	+++	false	true
		is dominant	false	false	
		is dominance-free	true	true	

This application of the two rules reveals their limitations. While DOMINANCE corresponds to the well-known stringent standard for decision-making, it is a rare occurrence for a single option to dominate all others. Such a criterion can often be overly restrictive. True, by adopting DOMINANCE, MOAC would close the moral gaps. But agents might frequently find themselves in scenarios without any morally right option. Thus, accepting DOMINANCE comes with the cost of accepting a lot of true moral dilemmas, i.e., such a version of MOAC has to give up all aspirations of having NO MORAL DILEMMAS. Consequently, we cannot hope that such a version will live up to PMH in any version. Further, note that *when there is a dominant option, we always can use CONDITIONALIZATION and apply the SURE-THING PRINCIPLE which, in combination, will always yield the same result as combining MOCOR with DOMINANCE.* We thus can ignore DOMINANCE for the remainder of this project.

DOMINANCE-free, on the other hand, ventures to the other end of the spectrum. This criterion implies a high degree of permissiveness. In interesting collective decision situations, like TROUBLEMAKERS are by design, several options won't be dominated by any other option. Consequently, under the DOMINANCE-free criterion, a plethora of actions might be concurrently deemed right and thus fail to offer the concrete and meaningful moral guidance that PMH asks for.

Fortunately, dominance-based methods are not the only non-aggregative amendments available. The toolbox of decision-making under uncertainty holds numerous other established methods, three of which are worth introduction in the context of this project.

### 7.3.2.2 MAXIMIN and MAXIMAX

Whoever mentions »dominance« has to mention »MAXIMIN« and »MAXIMAX«. These twin principles, foundational in decision and game theory, are often discussed in tandem due to their contrasting orientations toward risk. As we have seen, dominance-based amendments, in a sense, zero in on universally preferable options. On the other hand, when dominance doesn't provide clear guidance, which is often the case and certainly is so with TROUBLEMAKERS (as shown above), MAXIMIN and MAXIMAX (and a zoo of further principles) step in. They spotlight the strategic distinction between safeguarding against the worst-case scenarios and gunning for the best possible outcomes. Beyond theoretical significance, the MAXIMIN principle, in particular, has also gained a reputation in terms of ethical theorizing, primarily, of course, because of its place value in John Rawls' theory of justice (Rawls [1971]). Very roughly, Rawls advocated for a society where inequali-

ties are structured such that the least advantaged benefit the most, hence advocating a **MAXIMIN** approach in the realm of distributive justice.

As always, we start with rigorous definitions:

**Definition 7.8 (MAXIMIN)** *Let  $D$  be a collective decision situation with an actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be a valuation function.*

*An option  $\phi \in \Phi_A$  in  $\Gamma_D$  is right for an agent  $A \in \mathcal{A}_D$  in  $C$  if and only if the worst final outcome associated with  $\phi$ , given  $C$ , is at least as good as the worst final outcome associated with any other option  $\phi'$  available to  $A$  in light of  $C$ .*

*Formally:*

$$\exists \phi' \in \Phi_A : \min_{\gamma' \in \Psi_{\phi'}} \text{Val}(\gamma') > \min_{\gamma \in \Psi_\phi} \text{Val}(\gamma).$$

**Definition 7.9 (MAXIMAX)** *Let  $D$  be a collective decision situation with an actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be a valuation function.*

*An option  $\phi \in \Phi_A$  in  $\Gamma_D$  is right for an agent  $A \in \mathcal{A}_D$  in  $C$  if and only if the best possible final outcome associated with  $\phi$ , given  $C$ , is at least as good as the best possible final outcomes associated with any other option available to  $A$  in light of given  $C$ . Formally:*

$$\exists \phi' \in \Phi_A : \max_{\gamma' \in \Psi_{\phi'}} \text{Val}(\gamma') > \max_{\gamma \in \Psi_\phi} \text{Val}(\gamma).$$

Here is, once again, an application of the principles to **TROUBLEMAKER** with the default structure:

		Y		min	max
		$\phi_Y$	$\neg\phi_Y$		
$\times$	$\phi_X$	—	---	---	—
	$\neg\phi_X$	---	+++	---	+++
min		---	---		
max		—	+++		

Thus, according to MAXIMIN, every option is right, while according to MAXIMAX, only  $\neg\phi$  is right. It turns out that, rather unsurprisingly, MAXIMAX is extensionally equivalent to MAXIMIZATION. MAXIMIN, on the other hand, is very permissive and will normally not guarantee the best outcome if every agent acts in accordance with the resulting assessments.<sup>135</sup>

Before we can turn to the final evaluation of amendments, there remains one variant that is worth introducing.

### 7.3.2.3 MIXED STRATEGIES

There is another method that, again, requires us to enter the domain of probabilistic decision-making, even though it does not come with objective probabilities as metaphysical baggage: The idea of going for *mixed strategies*. Rather than committing to a single action, such approaches allow agents to act according to probability distributions over their options, asking them to opti-

<sup>135</sup>Both can be refined by extending them to their lexicographic versions. For instance, we can extend MAXIMIN (MAXIMAX) such that we can rank options with equally good best minima (maxima) with respect to the second best (worst) outcome and so on, recursively, until only combinations with identical evaluative profiles are ranked equally high. I spare the details here and leave them for another occasion. We later revisit this option briefly, though.

mize the expected outcome. Effectively, this corresponds to adding infinitely many new potential options into the agent's option spaces, as agents can mix actions in countless ways based on different probability assignments. This approach provides a balanced blend of risk and reward, paving the way for more flexible and adaptable decision-making frameworks.

However, mixed strategies, at first glance, might appear to complicate matters and to modify the collective decision situations unjustifiedly. However, this enrichment of option space offers several compelling advantages. First, it may do justice to how we make real decisions. In any case, given my introspection, it does not seem at all far-fetched to me that if I had to make certain decisions again in certain situations, I would *not necessarily make the same* decisions. And not because I know things now that I didn't know then. Even if I assume I would be in exactly the same (epistemic, emotional, etc.) state, this observation seems plausible. Sometimes, I claim, there seem to be several options on the table, and which one we end up realizing seems, in a sense, to be a matter of chance. In that sense, I think it's possible that we *can* act according to such probability distributions – and equally, I think it's quite plausible that maybe we *should* do so sometimes.

Furthermore, introducing mixed strategies has proven theoretically fruitful in many fields as it facilitates a deeper and more comprehensive equilibrium analysis, often revealing Nash equilibria<sup>136</sup> where no >optimal strategy< were apparent in a purely non-probabilistic setting. In other words, by lever-

---

<sup>136</sup>A Nash equilibrium, named after John Nash (Nash 1950) Nash (1951), roughly, refers to an outcome in a game (i.e., a collective decision situation) where each player's strategy is optimal given the strategies chosen by the other players. In other words, given the choices of the other agents, no agent has a unilateral incentive to deviate from their chosen strategy.

In a sense, we encountered such equilibria long ago: Non-global equilibria are what made TROUBLEMAKER challenging (cf. Section 6.4.1)

aging the tools of probability and expected utility theory, mixed strategies enable agents to optimize outcomes in ways that wouldn't be possible with a limited set of pure strategies. In essence, while the approach might seem counterintuitive, it offers a richer, more reflective, and more analytically powerful framework for understanding decision-making in complex scenarios.

Finally, inflating the option spaces to infinity does *not* cause any formal challenges. The available mathematical toolbox can handle them well. So, although I acknowledge that some burden comes with accepting the possibility of mixed strategies, I think this burden weighs much less than that which would come with EXPECTED UTILITY. After all, the probabilities introduced are not metaphysical but much more analytical.

Besides that, I am not the first who had this idea. Actually, it is quite old. Recall that in the essay on utilitarianism (Smart 1973), Bernard Williams introduced the example of JOB MARKET, which was later employed by Jonathan Glover (Glover and Scott-Taggart 1975) to ignite the debate on the CHALLENGE (cf. Section 3.4 on page 75). However, often overshadowed in discussions of this essay is J.C.C. Smart's intriguing proposal to expand MOAC into collective contexts through the introduction of mixed strategies in response to Richard Brandt's Frenchman case (cf. Section 3.4):

Suppose that, in wartime England, people are requested, as a measure essential for the war effort, to conserve electricity and gas by having a maximum temperature of 50 degrees F. in their homes. A utilitarian Frenchman living in England at the time, however, argues as follows: »All the good moral British obviously will pay scrupulous attention to conforming with this request. The war effort is sure not to suffer from a shortage of electricity and gas. Now, it will make no difference to the war effort whether I personally use a bit more gas, but it will make a great deal of difference to my comfort. So, since the public welfare will

be maximized by my using gas to keep the temperature up to 70 degrees F. in my home, it is my duty to use the gas.«

According to the act-utilitarian theory, this argument is perfectly valid. But we should not take it seriously in fact. Why not? At least part of the reason is that we think that, if a sacrifice has to be made for the public good, all should share in it equally. Imagine the outcry in Britain, if it became known that members of the Cabinet, who knew that electricity and gas were in good supply because of the country's willingness to sacrifice, used this argument to justify using whatever power was necessary to keep their homes comfortable.

Smart's response was to advocate mixed strategies (Smart [1973], p. 59):

There is a circularity in the situation which cries out for the technique of game theory.

There are three types of possibility: (a) he can decide to obey the government's request; (b) he can decide not to obey the government's request; (c) he can decide to give himself a certain probability of not obeying the government's request, e.g. by deciding to throw dice and disobey the government's request if and only if he got a certain number of successive sixes.

To decide to do something of type (c) is to adopt what in game theory is called »a mixed strategy«. On plausible assumptions it would turn out that the best result would be attained if each member of the act-utilitarian society were to give himself a very small probability  $p$  of disobeying the government's request.

I think this idea is plausible and worth spelling out in a bit more detail than Smart did. For this, we first define, given a collective decision situation  $D$  be a with actual context  $C$ , domain  $\Psi$  and some valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$ :

$$EV(\Pr_\Psi) := \sum_{\Upsilon \in \Psi} \Pr_\Psi(\Upsilon) \text{Out}_C(\Upsilon)$$

So what we are looking for are a selection of probability distributions  $\Pr_{\Phi_A} \in \Pr(\Phi_A)$  for all agents  $A \in \mathcal{A}_D$  (where  $\Pr(\Phi_A)$  denotes the set of all possible probability distributions over  $\Phi_A$ ) that in combination maximize  $EV(\Pr_\Psi)$ .

This naturally raises the question of how exactly the probability distribution  $\Pr_\Psi$  is related to the sought probability distributions  $\Pr_{\Phi_A}$ . Since I have narrowed down the investigation to maximal COORDINATION CASES a strong form of INDEPENDENCY OF ACTION is guaranteed (cf. Section 5.2.2). Thus, we can get a particularly simple answer to this question. First, note that  $\Pr_\Psi$  is simply the joint probability over the actions of the agents involved in the combination, i.e., for a  $\Upsilon$  of  $n$  actions we have:

$$\Pr_\Psi(\Upsilon) = \Pr_\Psi(\langle \phi_1, \dots, \phi_n \rangle) := \Pr_\Gamma(\phi_1, \dots, \phi_n).$$

Second, in the case of independent events,  $\Pr_\Gamma(\phi_1, \dots, \phi_n)$  can be simply decomposed into a simple factorization, i.e., into a product of the marginal probability distributions. Let us agree on the notion that  $\Upsilon_A$  denotes the option  $\phi \in \Phi_A$  with  $\phi \in \Upsilon$  and, to ensure that this notion does not run empty, we define  $\mathcal{A}_\Upsilon$  to denote the subset of  $\mathcal{A}$  only containing agents that contribute to  $\Upsilon$  (and thus,  $\Upsilon_A$  is defined for every  $A \in \mathcal{A}_\Upsilon$  for arbitrary  $\Upsilon \in \Psi$  per definition).

Then we have<sup>137</sup>

$$\Pr_\Psi(\phi_1, \dots, \phi_n) = \prod_{i=1}^n \Pr_{\Phi_{A_{j_i}}}(\phi_i).$$

Which simplifies now to the following equation with much less index complexity:

$$\Pr_\Psi(\Upsilon) := \prod_{A \in \mathcal{A}_\Upsilon} \Pr_{\Phi_A}(\Upsilon_A).$$

We now can define:

---

<sup>137</sup>Here  $\{j_1, \dots, j_n\}$  is an index set encoding the order of actions by the agents enumerated accordingly.

**Definition 7.10 (MIXED STRATEGIES)** Let  $D$  be a collective decision situation with actual context  $C$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.

To perform an option  $\phi \in \Phi_A$  in  $\Gamma_D$  with probability  $\Pr_{\Phi_A}(\phi)$  is right for an agent  $A \in \mathcal{A}_D$  (in context  $X$ ) if and only if  $\Pr_{\Psi} \in \arg \max_{\Pr' \in \mathcal{P}_{\Psi}} EV(\Pr'_{\Psi})$ .

It is easy to calculate the searched  $\Pr_{\Psi}$  since we know that

$$EV(\Pr_{\Psi}) := \sum_{\Upsilon \in \Psi} \Pr_{\Psi}(\Upsilon) \text{Out}_C(\Upsilon)$$

Thus, we can just calculate the distribution  $\Pr_{\Psi}$  such that the derivative vanishes, i.e.,  $\frac{\partial EV(\Pr_{\Psi})}{\partial \Pr_{\Psi}} = 0$ . In consideration of the function values at the boundary (i.e., for performing one option with a probability of one), we can thus identify the distribution  $\Pr_{\Psi}$  that maximizes  $EV(\cdot)$ .

In any case, this approach, again, makes it necessary to adjust the criterion of rightness to make room for expectation values. Here is a suggestion:

**Principle 7.1 (MOCOR (EU))** Let  $D$  be a collective decision situation of agents  $\mathcal{A}$  with domain  $\Psi$  and actual context  $C$ .

$\phi \in \Phi_A$  is right for agent  $A \in \mathcal{A}$  (given  $C$ ) if and only if  $A$ 's choice followed a probability distribution  $\Pr_{\Phi_A} : \Phi_A \rightarrow [0, 1]$  where  $\Pr_{\Phi_A}$  belongs to a set of probability distributions of all agents such that  $\Pr_{\Psi} : \Psi \rightarrow [0, 1]$  defined as

$$\Pr_{\Psi}(\Upsilon) := \prod_{A \in \mathcal{A}_{\Upsilon}} \Pr_{\Phi_A}(\Upsilon_A)$$

maximize  $EV(\Pr_{\Psi})$ , relative to  $C$ .

This necessity to touch MOCOR is not a knockout criterion, but the willingness to go this road requires justification. This brings us to the final question of what remains to be done.

Amendment	Remarks	Still in the Game?
SUMMATION	promising	in
MAXIMIZATION	promising	in
EXPECTED UTILITY	Although in principle it is not impossible to do something with it, the assumption of objective probabilities brings too great a burden with it. It might be a bigger project to fill this approach with life. Put aside for the time being.	out
DOMINANCE	Allows moral dilemmas and otherwise offers nothing that we could not derive long ago with the SURE-THING PRINCIPLE.	out
DOMINANCE-free	Too permissive especially in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MAXIMIN	To permissive in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MAXIMAX	Extensionally equivalent to MAXIMIZATION but comes with modification of MOCoR	out
MIXED STRATEGIES	promising	in

**Table 7.1:** The introduced amendments, which ones are still in the game, and which ones are no longer (and, in brief, why the amendments that are no longer in the game are not so anymore).

## 7.4 What Remains to Be Done

Up to now, we've achieved several insights. First, we have seen, how the APPROACH brought genuine consequences in the form of reduced decision situations of individual options to the consequentialist workbench, even in case of the most complex collective decision situations. Second, we have observed that various amendments would, in principle, allow the consequentialists to employ these newly discovered consequences in order to fill the gaps that imply the REAL CHALLENGE. While several of these amendments could be set aside due to various considerations, others remain viable contenders for addressing the REAL CHALLENGE in an overall promising way.

An overview can be found in Table 7.1

But which of these amendments should MOAC adopt to address not only the REAL CHALLENGE, but also simultaneously the CHALLENGE? After all, we do not just fill the gaps, but we want to do it in a way that does not re-raise the CHALLENGE. In light of the PYRAMID, it seems natural to focus on the CHALLENGE<sub>int</sub> first and revisit PMH, hoping that it defines MOAC's true collective objective. This would allow us to identify those states in a GEF that represent those final outcomes to which MOAC should lead the way so that, in turn, we can then select an amendment that closes the gaps accordingly. So, before we decide on an amendment, it's time to explore the legitimacy and limits of PMH.

# **Chapter 8**

## **On Reasonable Disharmonies and The Quest for The Best Amendment**

At this point, we have a clear picture in mind: We have established good reasons to believe that collective decision situations, in the spirit of COMPOSITIONALISM, are nothing more than a collection of individual decision situations. The consequences necessary for this, long neglected by Camp MOAC, have been provided by the APPROACH. In addition, we have found several theoretical supplements, which I called »amendments«, which enable the champions MOAC's to close the gaps causing the REAL CHALLENGE. We were able to reject some of the amendment candidates at this stage. The remaining candidates will be tested for suitability in this chapter.

The question, of course, is *how* these tests could look like and how, in the end, shall be decided between the different amendments. The answer obviously arises from the larger context of this project: We need a solution to the REAL CHALLENGE, and one that also solves the CHALLENGE in its various versions.

In light of the PYRAMID, it seems natural to focus on the CHALLENGE<sub>int</sub> first and revisit PMH, hoping it defines MOAC's ›true‹ collective objective. This would allow us to identify paths through the GEF towards the final outcomes to which MOAC should lead the way so that, in turn, we can select an amendment that closes the gaps accordingly. So, before we decide on an amendment, we must explore the legitimacy and limits of PMH.

The task of this chapter, thus, is threefold. First, we revisit PMH and look for guidance with respect to this question. We will realize quickly that MH and the other traditional explications of PMH are actually too strong as they are incompatible with even more fundamental objective-consequentialist commitments. At the end of this investigation, we will have a suitably hedged but rather vague version of MH. The second goal of this chapter, then, is to fill that formulation with life and to make it more precise. I introduce the otherwise well-established concept of *policies* to the consequentialist toolbox and investigate how policies can be ranked from MOAC's perspective, specifically in light of our new version of MH. Finally, this allows us to revisit and evaluate the formerly introduced amendments since MOAC, together with each of these amendments, induces a policy that we can now assess. Thus, at the end of this chapter, I finally can advocate a specific amendment and argue how we have conquered the PYRAMID with one single strike.

## 8.1 Revisiting PMH

In the introduction, I motivated the CHALLENGE with recourse to

**View 1.1 (CONGRUENCE)** *The right and the best are congruent in the sense that doing what is right goes (necessarily) hand in hand with bringing about the morally best consequences that can be brought about.*

Based on this core conviction shared by MOAC, I proclaimed two readings, one with individual and one with collective character. The individual one is that the right action brings about the best possible consequences. This reading is so dear to MOAC theories that it is captured in what makes them MOAC theories, i.e., the criterion of rightness they share, namely:

**Principle 2.2 (MOC or)** *Let  $D$  be an individual decision situation involving an agent  $A$  with a set of options  $\Phi$  and with actual context  $C$ . An action  $\phi \in \Phi$  is right for  $A$  if and only if, relative to  $C$ , there is no alternative action  $\phi' \in \Phi$  with better consequences than  $\phi$ .*

The collective reading I called the PRINCIPLE OF MORAL HARMONY. The specific formulation I distilled from several sources was this principle:

**Principle 5.1 (COLLECTIVELY MAXIMIZING)** *Let  $D$  be a collective decision situation with domain  $\Psi$  and with actual context  $C$ . If  $\Upsilon \in \Psi$  consists only of right actions, then there is no (and cannot be) alternative  $\Upsilon' \in \Psi$  with better consequences than  $\Upsilon$  relative to  $C$ .*

As soon as you broaden your view and look at cases other than the typical TROUBLEMAKERS, it becomes clear that something fundamental is shady with that principle: As plausible as COLLECTIVELY MAXIMIZING may sound at first, it *cannot* be true on closer inspection – at least not in that generality. In the following, I will first explicate that issue with PMH and argue that the necessary cuts and restrictions that come with this insight are greater than is sometimes acknowledged implicitly. Subsequently, I will propose a modified formulation of MH, which, although underdetermined in certain respects, is an excellent benchmark for our remaining gap-filling project.

### 8.1.1 The Limits of PMH

Recall Parfit's contour maps from the last chapter (as a reminder, see Figure 8.1) that we might understand as a taxonomy of COORDINATION CASES. Let us have a closer look at what Parfit calls »Symmetrical Coordination Cases« (which I think is a misnomer, because all the cases that Parfit lists there are symmetric, cf. Section 5.2.1.3). Here is a modified version of Regan's WHIFF AND POOF with a fitting normal form:

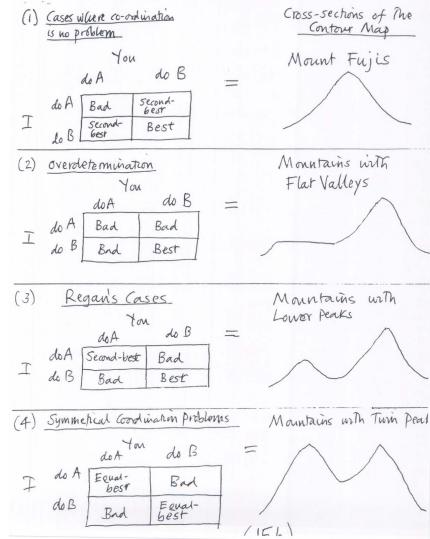


Figure 8.1: Parfit's contour maps (again).

		Poof	
		not-push	push
Whiff	not-push	10	0
	push	0	10

In such a case, to assume that morality could illuminate the way for the two agents to reach one of the best possible worlds seems tantamount to believing in magic. The combination of symmetry and non-uniqueness of the optimum brings such a belief into serious trouble. To get to the heart of the problem, the following consideration helps us. Suppose  $T$  was a moral theory satisfying COLLECTIVELY MAXIMIZING. Then it must either be right for Whiff to push his button and wrong for Poof to push his button; or it must be just the other way around, i.e., right for Poof to push and wrong for Whiff not to push. But what is the morally relevant difference supposed to be? Even without

adopting APP PROACH, it seems difficult from a consequentialist point of view to point to a relevant difference. The simply inverted rows and columns are too similar.  $T$  would have to magically single out one of the optima and then assess the actions as right that, in combination, produce that chosen optimum. But why choose the one optimum over the other?

However, the problem goes deeper than merely being an implausible requirement for cases with multiple global optima because we can construct cases in which satisfaction of COLLECTIVELY MAXIMIZING is not only implausible but would even be *logically incompatible with objective consequentialism*. Consider the following case, illustrated in Figure 8.2:

**Case 8.1 (Seaman Clumsy)** *Amid the vast expanse of the high seas, three sailors are aboard a ship: the navigator Mikel, the steerswoman Laika, and a seaman known as Clumsy. As the ship sails through a violent hailstorm, Clumsy, who is a notably poor swimmer, suddenly topples overboard.*

*Both Mikel and Laika, positioned on opposite ends of the vessel, witness Clumsy's peril. The storm's intensity makes it impossible for them to ascertain if the other has seen the incident, and the raging elements prevent any form of communication or coordination. Each of them faces an immediate decision: they can either dive into the turbulent waters to aid Clumsy or turn around to retrieve the lifebuoy to toss to him.*

*The ship's high bulwark presents a significant challenge. If both Mikel and Laika decide to jump, they'll find it impossible to hoist themselves back onto the ship. Consequently, all three sailors would be doomed to the depths. Conversely, if both decide to fetch the lifebuoy, the time it would take amid the storm would be too long, and Clumsy would succumb to the sea before they can assist.*

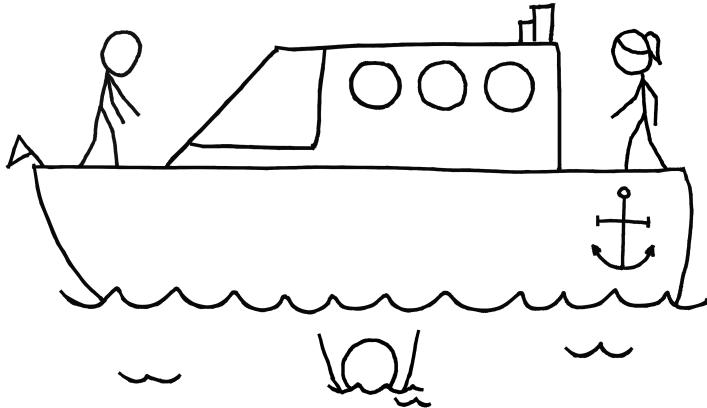


Figure 8.2: The situation from SEAMAN CLUMSY.

The optimal scenario, no doubt, would be for one of them to dive in to help Clumsy stay afloat while the other retrieves the lifebuoy. However, if neither of them jumps, Clumsy dies; and if both jump, they all plunge. Thus, we may represent SEAMAN CLUMSY in the following normal form:

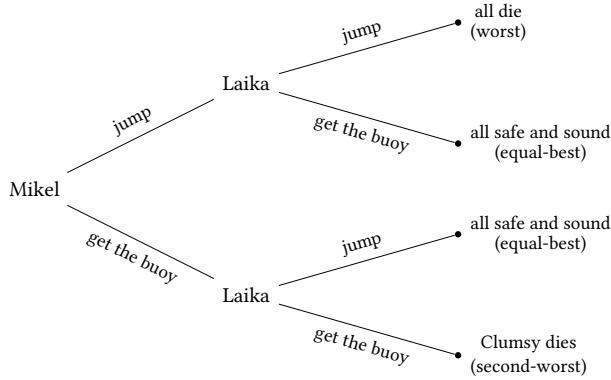
		Mikel	
		jump	get the buoy
Laika	jump	worst	equal-best
	get the buoy	equal-best	second-worst

SEAMAN CLUMSY is not only a symmetric COORDINATION CASE, but also an interesting one, i.e., the only assessments we get, according to the traditional analysis we reconstructed, are conditional assessments, i.e.,

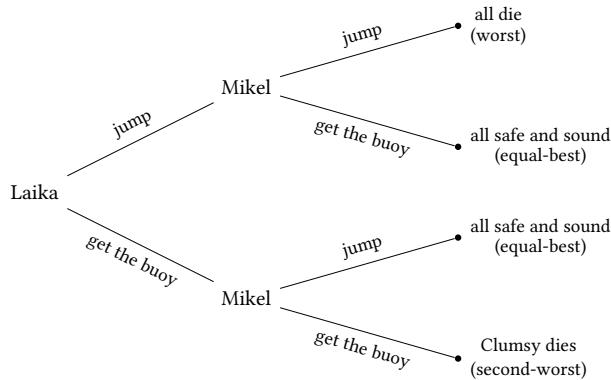
(48) If Mikel (Laika) jumps, it is right for Laika (Mikel) to get the buoy.

(49) If Mikel (Laika) gets the buoy, it is right for Laika (Mikel) to jump.

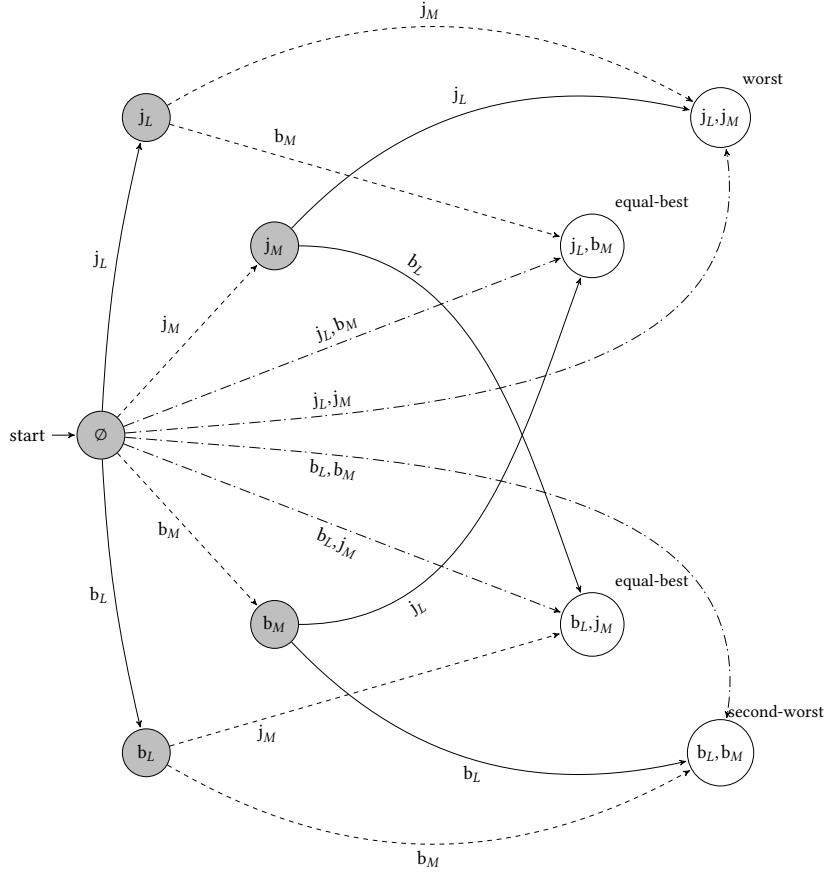
However, if we adopt APPROACH, we find that both Mikel and Laika are in valuative identical decision situations, namely



and



Alternatively, we can make use of our new representational vehicle, the GEF. Figure 8.3 shows **SEAMAN CLUMSY**'s GEF. We can instantly see that whoever acts first brings the second acting agent in *valuatively identical* decision situations. If Laika (Mikel), as the first-acting agent, jumps, then Mikel (Laika) has the choice between bringing about one of the equal-best worlds (by getting the buoy) or bringing about the worst world (by jumping). Vice versa, if Laika (Mikel), as the first-acting agent, gets the buoy, then Mikel (Laika) has the choice between bringing about one of the equal-best worlds (by jumping) or bringing about the second-worst world (by getting the buoy). As a result, the following is guaranteed: No matter what option (or options)

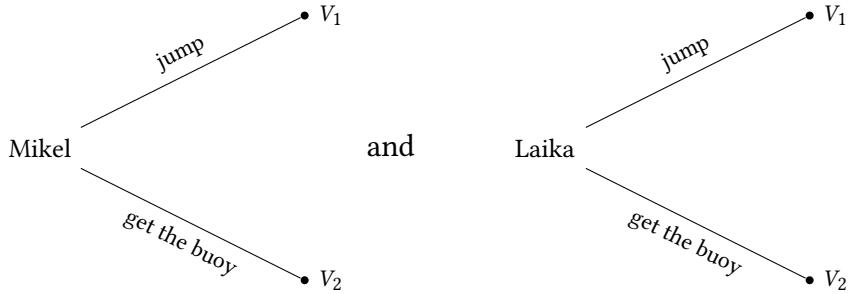


**Figure 8.3:** Collective Extensive Form of SEAMAN CLUMSY (where  $j$  stands for jumping,  $b$  for getting the buoy,  $M$  for Mikel and  $L$  for Laika).

MOAC is going to assess as right for the first-acting agent, it should, in a sense, be the same kind of option(s) for both agents.

Therefore, to make the argument that for such cases, objective consequentialists are actually forbidden to embrace COLLECTIVELY MAXIMIZING, it *does not matter* how exactly we assess the respective individual decision situations between which the agents have to choose. As long as the assessment is consequentialist, i.e., it depends solely on what outcomes they can choose between and not, say, on which agent is in the situation, we can say that the respective decision situations must be equally evaluated or ranked. Let's say, for instance, the value is  $V_1$  for the upper action situation (in which

a jump would mean the death of all involved) and  $V_2$  for the lower action situation (in which a jump would ensure the survival of all). Then we have these situations in front of us:



It should be obvious that no objective act-consequentialist theories can come to different assessments in these two cases. This can also be explicitly shown. Recall:

**Definition 2.3 (Objective Consequentialist Theory (formal))** *T is an objective consequentialist theory if and only if it embraces an axiological sub-theory  $T_{Ax}$  with a valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and an objective consequentialist criterion of rightness  $T_{COR}$  such that, for all decision situations  $D \in \mathcal{I}$  and for all  $\phi \in \Phi_D : D, C \models_T R\phi$  if and only if  $T_{COR}(\phi)$ .*

*A criterion of rightness  $T_{COR}$  is objective consequentialist if and only if, for all  $D \in \mathcal{I}$  with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  (with D's actual context C)  $T_{COR}$  corresponds to a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :*

$$D, C \models_T R\phi \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))).$$

Indeed, this definition *implies* that any objectively consequentialist theory has another property that I will call *normative supervenience*.<sup>138</sup> Roughly

<sup>138</sup>The thoughts presented here are my own. However, I later found out that Krister Bykvist had the same insight more than twenty years ago, cf. Bykvist [2002] Bykvist [2003]. The similarities go so far that he also used the term »Normative Supervenience« (this is even how I found his article) in this context, even though the properties presented here are rather what he calls »Consequentialist Supervenience«.

speaking, for a theory to have this property means to be such that, in valuatively identical decision situations, the theory necessarily assigns the same moral status to corresponding options. We can explicate the involved terms.<sup>[139]</sup>

We first define:

**Definition 8.1 (Valuative Embedding)** *Let  $D, D' \in \mathcal{D}$  be two individual decision situations with*

$$D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$$

*and*

$$D' := \langle A', \Phi', \text{Out}'_{C'} : \Phi' \rightarrow \mathcal{O} \rangle$$

*and let  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  be a valuation function.*

*D can be valuatively embedded in  $D'$  (relative to  $\text{Val}$ ) if and only if there is a injection  $f : \Phi \rightarrow \Phi'$  such that for all  $\phi \in \Phi$ :*

$$\text{Val}(\text{Out}_C(\phi)) = \text{Val}(\text{Out}'_{C'}(f(\phi)))$$

Next, we define:

**Definition 8.2 (Valuative Identity)** *Let  $D, D' \in \mathcal{D}$  be two individual decision situations with*

$$D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$$

*and*

$$D' := \langle A', \Phi', \text{Out}'_{C'} : \Phi' \rightarrow \mathcal{O} \rangle$$

*and let  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  be a valuation function.*

*D is valuative identical to  $D'$  (relative to  $\text{Val}$ ) if and only if D can be valuatively embedded in  $D'$  and vice versa.*

---

<sup>139</sup>This is the moment promised in Section 3.5.2.1: We now have to precisely define the then intuitively captured notion of structural equivalence.

This is to say that for every option in the one decision situation, there is an option with an equally good outcome in the other decision situation. Thus, for valuatively identical decision situations, we have an injection from the one decision situation's options space into the other's and vice versa, and so we know that there is a bijection between these option spaces. We call such a bijection a »*correspondence relation*«.

Based on these notions, we can define

**Property 8.1 (NORMATIVE SUPERVENIENCE)** *Let  $D, D' \in \mathcal{D}$  be two valuative identical decision situations (relative to valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and correspondence mapping  $f : \Phi_D \rightarrow \Phi_{D'}$ ). For every  $\phi \in \Phi_D$  and  $\phi' \in \Phi_{D'}$ : If  $f(\phi) = \phi'$ , then  $\phi$  and  $\phi'$  have the same moral status in their respective decision situation.*

It is indeed easy to see that every objective consequentialist moral theory (in accordance with Definition 2.3) has NORMATIVE SUPERVENIENCE.<sup>140</sup>

---

<sup>140</sup>The proof is easy, but a bit lengthy: Let  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  and  $D' := \langle A', \Phi', \text{Out}'_{C'} : \Phi' \rightarrow \mathcal{O} \rangle$  (where  $C$  is the actual context of  $D$  and  $C'$  the actual context of  $D'$ ) be the representations of two valuative identical decision situations (relative to valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and correspondence mapping  $f : \Phi_D \rightarrow \Phi_{D'}$ ). To prove the claim, we need to show that for every  $\phi \in \Phi_D$  and  $\phi' \in \Phi_{D'}$ : If  $f(\phi) = \phi'$ , then

$$D, C \models_T R\phi \text{ if and only if } D', C' \models_T R\phi'.$$

Since  $T$  falls under Definition 2.3, there is a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D,C})}$  such that for all  $\phi \in \Phi$ :

$$D, C \models_T R\phi \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi)))$$

and there is a predicate  $\chi_{T, \text{Val}(\mathcal{O}_{D',C'})}$  such that for all  $\phi' \in \Phi'$ :

$$D', C' \models_T R\phi' \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D',C'})}(\text{Val}(\text{Out}'_{C'}(\phi'))).$$

Thus, it suffices to show that, if  $f(\phi) = \phi'$ , then

$$\chi_{T, \text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))) \text{ if and only if } \chi_{T, \text{Val}(\mathcal{O}_{D',C'})}(\text{Val}(\text{Out}'_{C'}(\phi')))$$

Now, since  $D$  and  $D'$  are valuative identical decision situations (relative to valuation function  $\text{Val} : \mathcal{W} \rightarrow \mathcal{V}$  and correspondence mapping  $f : \Phi_D \rightarrow \Phi_{D'}$ ), we know several things that help us to establish this equivalence. First, we know that for all  $\phi \in \Phi$ :  $\text{Val}(\text{Out}_C(\phi)) = \text{Val}(\text{Out}'_{C'}(f(\phi)))$ . Thus, since, by assumption,  $f(\phi) = \phi'$ , we know that  $\text{Val}(\text{Out}_C(\phi)) = \text{Val}(\text{Out}'_{C'}(\phi'))$ . Second, we know that  $\text{Val}(\mathcal{O}_{D,C}) = \text{Val}(\mathcal{O}_{D',C'})$  because we defined

$$\text{Val}(\mathcal{O}_{D,C}) := \{ \text{Val}(O) \in \mathcal{V} \mid O \in \mathcal{O}_{D,C} \},$$

This brings us back to *SEAMAN CLUMSY*. Maika and Laika's decision situations are evaluative identical since we can map Mikel's option to jump to Laika's and his option to get the buoy to her option to get the buoy. Thus, since MOAC is an objective consequentialist theory – of which we have long since assured ourselves in Chapter 2 –, whatever moral status MOAC assigns to any of this option it has to assign for the other. Given the standard assumptions with respect to the values  $V_1$  and  $V_2$ , most importantly, the assumption that they are comparable and, thus, that we can rank them, we have four obvious possible assessments to consider:

- Both options – to jump and to get buoy – are meant to be right according to an improved version MOAC. In this case, it cannot be guaranteed that Mikel and Laika, when acting in accordance with MOAC, bring about the best outcome that they together could bring about. It could well be that the one jumps and that the other gets the buoy; but they even could both get the buoy or jump. The latter combinations are suboptimal and one even results in the *worst* possible outcome, namely the death of all three.
- Alternatively, such a version of MOAC might assess only the option to jump as being right. Then Mikel and Laika, if they acted according

---

which we can rewrite (given that, by definition,  $\mathcal{O}_C = \{ \text{Out}_C(\phi) \mid \phi \in \Phi \}$  for an individual decision function  $D$  with option space  $\Phi$  and outcome function  $\text{Out}_C$ ) as

$$\text{Val}(\mathcal{O}_{D,C}) = \{ \text{Val}(\text{Out}_C(\phi)) \in \mathcal{V} \mid \phi \in \Phi \}$$

which, in light of the above first established identities and the existence of bijection  $f$ , we can rewrite to the set  $\{ \text{Val}(\text{Out}'_{C'}(\phi')) \in \mathcal{V} \mid \phi' \in \Phi' \}$ . Thus,  $\text{Val}(\mathcal{O}_{D,C}) = \text{Val}(\mathcal{O}'_{D',C'})$ . Therefore, by substituting both, the parameter and the argument of  $\chi$ , it holds indeed that

$$\chi_{T,\text{Val}(\mathcal{O}_{D,C})}(\text{Val}(\text{Out}_C(\phi))) \text{ if and only if } \chi_{T,\text{Val}(\mathcal{O}'_{D',C'})}(\text{Val}(\text{Out}'_{C'}(\phi'))).$$

In other words: If  $T$  is a objective consequentialist theory, then it adheres to NORMATIVE SUPERVENIENCE.

to MOAC, would not only *miss* the optimal results; they would even produce the *worst* possible result with certainty.

- Finally, such a version of MOAC could assess only the option to get the buoy as right. Then Mikel and Laika, if they acted accordingly, would miss the optimal result, but they would at least be guaranteed to produce the second-best result.
- In light of the lessons of the last chapter, there is a fourth possibility. Such a version of MOAC could command both to jump with a certain probability (greater than zero and lesser than one) and to fetch the buoy with the opposite probability. But even then, achieving the best result is just *possible* but not guaranteed. After all, the probability for the respective corresponding options must be equal (i.e., MOAC would need to assign the same probability distributions over the options of both agents, relative to correspondence mapping); that's what NORMATIVE SUPERVENIENCE requires. Therefore, even adopting such a *mixed* strategy cannot lead to an optimum with certainty.

My point here is this: even *without* having thought about how to assess or rank the individual decision situations that Mikel and Laika can produce in SEAMAN CLUMSY in concrete terms, we already know that *if* they are >reasonably< ranked (that is, in line with other traditional consequentialist assumptions and NORMATIVE SUPERVENIENCE), they are *guaranteed* to be assessed by MOCoR in a way that *allows for suboptimal* outcomes.

In other words, if we go along with APPROACH, we can solve the REAL CHALLENGE, but we still would not do justice to MH. However, this does *not* mean that APPROACH is problematic because it leads again into the CHAL-

LENGE for cases like SEAMAN CLUMSY (although it can possibly handle TROUBLEMAKERS well; but we will come to that later). Instead, it should be clear: To stick to PMH *in the form of* MH (or in the form of all the similar formulations we have encountered so far) for cases like SEAMAN CLUMSY *means to give up objective consequentialism*. After all, in these cases, as an objective-consequentialist theory, one must necessarily violate MH, since one is committed to NORMATIVE SUPERVENIENCE by logical implication. This is the takeaway of this section.

What we need, then, to seriously address the question of assessing the newly discovered consequences of actions in interesting collective decision situations is a *reasonable, acceptable explication of the general idea* PMH. This should then serve as our benchmark in filling in the gaps.

### 8.1.2 Upshot: Reasonable MORAL HARMONY

It should be clear by now that a straightforward formalization of the collective reading of CONGRUENCE, i.e., a naïve version of PMH like MH will be at odds with other, indisputable principles in certain cases. Thus, what is needed is a more informed formulation of PMH that accounts for further requirements of moral theories. Such a formulation has to avoid the violations of other well-justified or even theoretically entailed constraints – especially, it must not require the impossible from objective consequentialist theories. All this was anticipated already in the context of Broad's Property (cf. Section 3.5.1) and Pinkert's ON-THE-HOOK (cf. Section 3.5.2.2, particularly on page 121). So we are prepared and know that we need a kind of less-subjective version of their properties. Here is a proposal of a deliberately vague formulation:

**Principle 8.1 (COLLECTIVELY MAXIMIZING (Reasonable))**

*If all agents act (consistently) rightly, then they are guaranteed to produce the morally best outcome they can reasonably be expected to collectively bring about.*

This formulation must raise at least two questions. First, one should ask what it means in light of the OBJECTIVE VIEW that a particular outcome is to be expected. Shouldn't everything be set and determined? The answer, of course, is no. METHODOLOGICAL INDETERMINISM necessarily brings uncertainty into play as soon as multiple actions decide whether specific outcomes will occur. We learned this lesson in the last two chapters. The details, however, must still be worked out.

Second, we must ask what it means to expect an outcome *reasonably*. Obviously, this is meant to hedge the formulation against exceptions like the one SEAMAN CLUMSY-like cases or, more generally, cases with non-unique global optima raised for the applicability of PMH. The obvious question is, then, how to understand this hedge more *precisely*.

I will address these questions in the following section. In doing so, I will ultimately venture into the realms of formal decision theory where we find a rich and valuable repertoire of concepts, notions, and methods, which not only allows us to clarify the above formulation in both respects but also shows the way to fill the gaps posed by the REAL CHALLENGE without letting the CHALLENGE emerge again at all.

## 8.2 Amendments for Reasonable Pathfinding

The task is to figure out what exactly it means that the right action of all agents guarantees the morally best outcome that can reasonably be expected.

We proceed in two steps. First, we develop a precise understanding of what it means to act (consistently) right in this collective sense, i.e., what it means exactly that everyone acts rightly on all occasions. Here is the basic idea: Presuming the APPROACH, we consider the idea of a function  $\pi$  that, for a given individual decision situation  $D \in I$ , decides systematically for one of the available options available for the agent within that situation and performs corresponding actions. Such functions are well known in formal decision theory, especially in the context of Markov decision processes (MDPs) – formal models for sequential decisions under risk – and reinforcement learning, and are called *policies*. Policies allow to model the stringent following of rules – or probability distributions. In this respect, we can think of them as a formal but flexible vehicle of a moral theory, even though they have been predominately investigated in the domain of instrumental rationality. Making the concept of policies fruitful for this project's task and for (moral) consequentialists in general is the goal task of the following section.

Second, there is a need to develop a criterion for what makes a policy »reasonably defensible« in terms of an informed version of PMH – and what then makes the one reasonably defensible policy better than another in the eyes of MOAC. This gives us a method to evaluate policies, which we can then use in the second part of this chapter to evaluate various collective amendments. After all, as we will see, MOAC, in combination with any reasonable and defensible amendment, implies a reasonably defensible policy.

### 8.2.1 On Policies and Their Evaluation

In formal decision theory, particularly within the constructs of MDPs and prominently in reinforcement learning, the term »policy« possesses a specific

technical meaning. While at a cursory glance, the term might invoke a more general notion of guidelines or strategies, within this context, it's decidedly more precise. A policy, denoted typically by the symbol » $\pi$ «, is a function (based on some systematic, in practice often learned) procedure that, in a sense, determines an agent's action in every possible state, i.e., in arbitrary decision situations. Policies can be either deterministic or indeterministic. In the first case, these are functions  $\pi : \Phi \rightarrow \mathcal{W}$  that choose an option and perform the corresponding action (resulting in an outcome), while in the second case, these policies are probability distributions  $\pi : \Phi \rightarrow [0, 1]$ . A policy then specifies the probability with which an agent performs a certain action. Since probabilistic, i.e., indeterministic policies are a generalization of deterministic ones, *and* since one probabilistic amendment (**MIXED STRATEGIES**) is still in the game, I opt for probabilistic policies in the following.

The objective in many MDPs and reinforcement learning scenarios is to identify an *optimal* policy (typically called  $\pi^*$ ), i.e., a policy that maximizes the agent's *expected* reward over time. Analogously, one might think of a moral consequentialist policy as a guideline for consistent moral right-doing that directs an agent's actions in each relevant moral scenario to maximize the expected moral value over time. The connection to *reasonable COLLECTIVELY MAXIMIZING* should then be pretty obvious: The idea is that *defensible* moral theories yield *reasonable* policies.

What we are looking for in this chapter and thus in the context of this project, then, is ultimately a policy that is *as good as possible* in the sense of MOAC, i.e., a policy that leads us to the morally best possible results if all agents were to follow it strictly. In the following, I introduce that concept carefully. We start by explicating the above thoughts:

**Definition 8.3 (Policies)** Let  $D$  be an individual decision situation with option space  $\Phi$ . A policy for the individual decision situation  $D$  is a probability distribution  $\pi_D : \Phi \rightarrow [0, 1]$  that assigns each  $\phi \in \Phi$  some probability and, if the outcome of such  $\phi$  is not final, then  $\pi_D$  implies a policy  $\pi_{D \downarrow \phi}$  for all the decision situations  $D \downarrow \phi$  in all the  $\mathcal{O}_D$

Let  $D$  be a collective decision situation of agents  $A \in \mathcal{A}$  with corresponding option spaces  $\Phi_A$ . Let  $\mathbb{I}_D := \{\mathsf{D}_A \mid A \in \mathcal{A}\}$  be the decomposition of  $D$  and let  $\mathcal{O}_D^A := \{D \downarrow \phi \mid \phi \in \Phi_A\}$  be the sets of outcomes per agent. A policy for the collective decision situation  $D$   $\pi_D$  is a probability distribution  $\pi_D : \mathbb{I}_D \rightarrow [0, 1]$  that implies a policy  $\pi_{D_A}$  for all the decision situations  $D \downarrow \phi$  in all the  $\mathcal{O}_D^A$ .

A policy for an individual decision situation, thus, is just an assignment of probabilities to the options available in that situation that also implies a policy for further decision situations downstream. A policy for a collective decision situation  $D$ , in contrast, is an assignment of probabilities to the individual decision situations  $D \in \mathbb{I}_D$  available at the current context, i.e., in principle, a decision who acts next with what probability, that, in addition, implies a policy for every remaining decision situation in all the  $\mathcal{O}_D^A$  resulting from that next action. (Note that in practice, at least in the context of this project, policies are given by a procedure and thus are applicable to arbitrary decision situations. Thus, the policies they apply for the downstream decision situations are actually the same policy.) In the following, I assume<sup>[141]</sup> that it is equally probable for every agent to act next, i.e., that it holds, for a collective

---

<sup>[141]</sup>However, one could also choose, alternatively, some bias in this distribution that may be implied by some kind of meta-amendments. I cannot think of a convincing reason for exploiting this additional degree of freedom, but I cannot prove that there is no such reason. It thus might well be that it is worth for camp MOAC to explore this path, even though I won't do it.

decision situation  $D$  with decomposition  $\mathbf{I}_D := \{\mathbf{D}_A \mid A \in \mathcal{A}\}$ , that:

$$\pi_D(\mathbf{D}_A) = \frac{1}{|\mathcal{A}|}$$

Before I turn to the question of how exactly policies are related to moral theories and what we are still missing beyond MOCoR plus amendments to derive policies from MOAC theories effectively, I will first clarify how to evaluate policies and then, second, who to rank policies. So, first we, once again, lift valuations functions, this time to the level of policies. The basic idea is that the value of a policy corresponds to the expected utility of following it. Thus, we define:

**Definition 8.4 ((Expected) Value of a Policy)** *Let  $D$  be a individual decision situation with  $\mathbf{D} := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$ , let  $\pi_D : \Phi \rightarrow [0, 1]$  be a policy for  $D$ , and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The (expected) value of a  $\pi_D$  relative to an individual decision situations  $D$ , in symbols  $\text{Val}_D(\pi_D)$ , is*

$$\text{Val}_D(\pi_D) := \sum_{\phi \in \Phi_D} \pi_D(\phi) \cdot \text{Val}_{D \downarrow \phi}(\pi_D).$$

*In the case that via such a reduction we arrive at a final outcome, we define:*

$$\text{Val}_{D \downarrow \phi}(\pi_D) := \text{Val}_D(\phi)$$

*Let  $D$  be a collective decision situation of agents  $A \in \mathcal{A}$  with corresponding option spaces  $\Phi_A$ , let  $\mathbf{I}_D := \{\mathbf{D}_A \mid A \in \mathcal{A}\}$  be the decomposition of  $D$ , let  $\pi_D$  be a policy for  $D$  (and thus a set of policies  $\pi_{D_A}$  for all the  $D_A \in \mathbf{I}_D$ ), and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The (expected) value of a  $\pi_D$  relative to a collective decision situations to  $D$  (with  $\pi_{D_A}$  denoting the policies for the decision situations  $D_A \in \mathbf{I}_D$ ), in symbols  $\text{Val}_D(\pi_D)$ , is*

$$\text{Val}_D(\pi_D) := \sum_{D_A \in \mathbf{I}_D} \pi_D(D_A) \cdot \text{Val}_{D_A}(\pi_{D_A}).$$

Thus, the value of a policy relative to an *individual* decision situation is the expected value that follows from adhering to that policy. The value of a policy relative to a *collective* decision situation is the sum over the (expected) values of the policies for the individual decision situations of the involved agents weighted by the probability that they are actually instantiated.

The definition of ranking of policies in consequentialist terms is then very straightforward:

**Definition 8.5 (Consequentialist Policy Ranking)** *Let  $D$  be an individual decision situation with option space  $\Phi$  and let  $\pi$  and  $\pi'$  be two policies for  $D$ .  $\pi$  is to be preferred in consequentialist terms to  $\pi'$  relative to individual decision situation  $D$  (in symbols,  $\pi >_D \pi'$ ) if and only if  $\text{Val}_D(\pi) > \text{Val}_D(\pi')$ .*

*Let  $\mathcal{D} \subset \mathcal{I}$  be a set of individual decision situations. and let  $\pi$  and  $\pi'$  be two policies for  $\mathcal{D}$ .  $\pi$  is to be preferred in consequentialist terms to  $\pi'$  relative to the set of individual decision situation  $\mathcal{D}$  (in symbols,  $\pi >_{\mathcal{D}} \pi'$ ) if and only if  $\text{Val}_{\mathcal{D}}(\pi) > \text{Val}_{\mathcal{D}}(\pi')$ .*

*Let  $D$  be a collective decision situation and let  $\pi$  and  $\pi'$  be two policies for  $D$ .  $\pi$  is to be preferred in consequentialist terms to  $\pi'$  relative to collective decision situation  $D$  (in symbols,  $\pi >_D \pi'$ ) if and only if  $\text{Val}_D(\pi) > \text{Val}_D(\pi')$ .*

Based on these notions, we can finally turn to the questions of how we can derive policies from moral theories and how this helps camp MOAC decide on the »right« amendments.

### 8.2.2 Evaluating Amendments

Policies do not fall out of the air, of course. In the field of morality, they should be closely related to moral theories. Indeed, we can see a policy as the

direct and immediate expression of a MOAC theory – at least if we add a last and minor theoretical puzzle piece. We establish this connection in two steps. First, we say that a policy  $\pi$  is *consistent with a criterion of rightness of a theory T* (or shorter: consistent with  $T$ ) if and only if  $\pi$  is a policy for  $T$ 's domain  $I_T$  and for all individual decision situations  $D \in I_T$  (with actual context  $C$ ) it holds for all  $\phi \in \Phi_D$  :

$$\pi(\phi) > 0 \text{ if and only if } T, D, C \vDash R(\phi).$$

That is, right options (according to  $T$ 's criterion of rightness) and only right options are possible according to  $\pi$ , i.e., are assigned non-zero probability. Agents who act in accordance with a policy consistent with  $T$  can, thus never violate  $T$  – or, conversely, will *necessarily always satisfy*  $T$ . Note that for classical, deterministic moral theories, the above condition boils down to

$$\pi(\phi) = 1 \text{ if and only if } T, D, C \vDash R(\phi)$$

if and only if exactly one option is right in  $D$  given  $C$  according to  $T$ .

But what if there is more than one right action? I take it that it is usually assumed that the agent in such a situation can *arbitrarily* choose from the set of right actions (cf. Section 2.2).<sup>142</sup>

This brings us, second, to yet another potential probabilistic element and that is often negligently overlooked in normative ethics. As announced, it is not to be accompanied by heavyweight, metaphysical determinations but to

---

<sup>142</sup>In fact, I believe that this is the point where instrumental rationality and morality can interact: After all, the agent is under no moral constraint with respect to their choice from the set of right actions, so why not let him choose according to his own interests? Since, in the relevant cases, the inclinations, interests, and dispositions of the agents are not made explicit, they cannot play a role here. Therefore, in the following, I simply assume that there is no preference order of the agent with regard to the permitted actions that were not already accounted for within the moral quality of the outcomes.

be of an analytical nature. It comes from a rather trivial enrichment of our concept of a full-fledged normative theory: A complete moral theory should offer not only a criterion of rightness (and, in the case of consequentialist theories, also an axiological module and a relevance stance, cf. Section 2.3) but also a *selection rule* that determines, for a set of options assessed as right according to its criterion of rightness, which one of them is selected.

Is there a selection rule that we can plausibly impute to MOAC theories? In a certain sense, we can describe the challenge we are facing here again as a reduction of an existing individual decision situation. This time, however, the number of agents is not reduced by the action of one of the agents (or an assumption about it), but the set of available options is reduced to only the right options according to a certain moral theory. Thus it should be immediately clear that according to a traditional view of moral theories, their propositional power is exhausted. Certain options have been assessed as right, and thus, there is nothing more to say morally about the remaining situation. The remaining situation is similar to a Buridan's ass-like scenario. To have an example in the collective setting, we can again look at Figure 8.1 (cf. page 332). Of the four types of COORDINATION CASES, the fourth is a such situation. Relative to all the amendments discussed so far (and that are still in play), the two agents here face equally good choices. Since we want theories with NO MORAL DILEMMAS (cf. Section 4.3.1), both actions should better be right then according to our preferred MOAC theory (for which we are still searching). What, then, may we assume about the action of two agents who always act rightly in the sense of MOAC?

However, I am interested in the analysis of MOAC with respect to abstract decision situations. Thus, we need a more *systematic* approach to the

choice between options that are equal in every respect (and we can assume that any preferences of the agents are already represented in the outcomes, etc.). I think it is thus fair to assume that MOAC should suggest breaking such ›deadlocks‹ by *fair randomization*.<sup>143</sup> This is, I assume that MOAC theories use a simple probabilistic selection rule that chooses each correct action with equal probability. In other words, I assume that MOAC’s selection rule implies for every decision situation  $D \in I_T$  (with actual context  $C$ ) the following probability distribution (where  $T$  is an arbitrary MOAC theory):

$$\Pr_D^T : \Phi_D \rightarrow [0, 1], \quad \Pr_D^T(\phi) = \begin{cases} \frac{1}{|T(D, C)|} & \text{if } \phi \in T(D, C) \\ 0 & \text{otherwise.} \end{cases}$$

This allows us to define:

**Definition 8.6 (Theory-Induced Policy)** *Let  $T$  be a normative theory. A Policy  $\pi_T$  is called the policy induced by  $T$  (also called the  $T$ -induced policy) if and only if  $\pi$  is consistent with  $T$  and  $\pi$  mirrors  $T$ ’s selection rule.*

For a MOAC theory  $T$ , then, we have that  $\pi_T$  is a  $T$ -induced theory if and only if for all  $D \in I_T$  for all  $\phi \in \Phi_D$  it holds that

$$\pi(\phi) > 0 \text{ if and only if } T, D \vDash R(\phi)$$

and

$$\pi(\phi) = \Pr_D^T(\phi).$$

---

<sup>143</sup>Again, this excludes, for example, that the agents’ dispositions unconnected to the value of the outcomes play a separate role in this. For example, assuming that it was right for Laika to jump and it was right for her to get the body in *SEAMAN CLUMSY*, it would be equally likely that Laika jumps into the water and that Laika gets the buoy. So Laika jumps with a probability of 50%, no matter how deep the jump would be or how much Laika would like to look like a hero, etc. However, the toolbox outlined below has everything one would need if one wanted to include such complicating factors and the resulting statistical biases. I ignore them mainly for the sake of readability.

This, then, is to say that we can freely switch between any arbitrary MOAC theory  $T$  and its induced policy, which allows us to finally rank amendments:

**Definition 8.7 (Consequentialist Amendment Ranking)** *Let  $T$  be a moral theory with normative gaps, and let  $\Delta_1$  and  $\Delta_2$  be two amendments that fill these gaps. Let  $T_1$  and  $T_2$  be the two moral theories resulting from  $T$  adopting  $\Delta_1$  and  $\Delta_2$ , respectively.*

*Let  $D$  be an individual decision situation with option space  $\Phi$ .  $\Delta_1$  is to be preferred in consequentialist terms to  $\Delta_2$  relative to the individual decision situation  $D$  (in symbols,  $\Delta_1 >_D \Delta_2$ ) if and only if  $\pi_{T_1} >_D \pi_{T_2}$ .*

*Let  $\mathcal{D} \subset \mathcal{I}$  be a set of individual decision situations.  $\Delta_1$  is to be preferred in consequentialist terms to  $\Delta_2$  relative to the set of individual decision situation  $\mathcal{D}$  (in symbols,  $\Delta_1 >_{\mathcal{D}} \Delta_2$ ) if and only if  $\pi_{T_1} >_{\mathcal{D}} \pi_{T_2}$ .*

*Let  $D$  be a collective decision situation.  $\Delta_1$  is to be preferred in consequentialist terms to  $\Delta_2$  relative to the collective decision situation  $D$  (in symbols,  $\Delta_1 >_D \Delta_2$ ) if and only if  $\pi_{T_1} >_D \pi_{T_2}$ .*

Based on these definitions, we can now rank different amendments relative to (sets of) decision situations. Thus, it only remains to execute this program for the amendments still in play and with respect to the decision situations most important in the context of this work.

### 8.3 The Final Evaluation

We now have everything on the workbench to decide on the remaining amendments. We have these very amendments as candidates (see again Table 8.1); we have an idea of how to assess them in light of their ›performance capabilities‹ in terms of the ranking of their implied policies (in interaction

Amendment	Remarks	Still in the Game?
SUMMATION	promising	in
MAXIMIZATION	promising	in
EXPECTED UTILITY	Although in principle it is not impossible to do something with it, the assumption of objective probabilities brings too great a burden with it. It might be a bigger project to fill this approach with life. Put aside for the time being.	out
DOMINANCE	Allows moral dilemmas and otherwise offers nothing that we could not derive long ago with the SURE-THING PRINCIPLE.	out
DOMINANCE-free	Too permissive especially in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MAXIMIN	To permissive in combination with PMH-driven, optima-demanding intuitions (and this is where we are coming from).	out
MAXIMAX	Extensionally equivalent to MAXIMIZATION but comes with modification of MOCoR	out
MIXED STRATEGIES	promising	in

**Table 8.1:** The introduced amendments, which ones are still in the game, and which ones are no longer (and, in brief, why the amendments that are no longer in the game are not so anymore).

with MOCoR and the selection rule just established); and we have discussed two different interesting and challenging types of COORDINATION CASE to serve as ›testbeds‹ (TROUBLEMAKERS and cases of the structure of SEAMAN CLUMSY). The result of an amendment assessment based on these building blocks cannot then, of course, claim generality: There may be other amendments that have not been considered in my investigation, and there may be other important types of COORDINATION CASES that would deserve consideration. But at least the CHALLENGE will be tackled (because we consider TROUBLEMAKERS), and, by considering SEAMAN CLUMSY-like cases, our findings on the scope of PMH (keyword: NORMATIVE SUPERVENIENCE) will also be accounted for. What is missing is the performance of an evaluation – and a final judgment. These correspond to the last steps of this project.

### 8.3.1 Defining a Testbed

The proposed methodology for assessing possible amendments for MOAC theories is based on assessing the policy resulting from MOCoR together with the respective amendment relative to an *evaluative background*, i.e., relative to a (set of) decision situations. Since the central goal of this work is to solve the CHALLENGE for MOAC, classical TROUBLEMAKERS (with the structure of WHIFF AND POOF or Two FACTORIES) should definitely belong to the evaluation background. Furthermore, cases like SEAMAN CLUMSY should be part of it, which should serve as dicing stones with respect to the performance on cases with more than one optimum and especially with respect to NORMATIVE SUPERVENIENCE.

Thus, we will consider two types of COORDINATION CASES,<sup>144</sup> defined through their evaluative profiles. The first type, denoted by trouble (with decomposition  $I_{\text{trouble}}$ ), is given by the structural normal form of Two FACTORIES-like cases (with  $v_{12} = v_{21} < v_{11} < v_{22}$ ), recall:

trouble	Y	
	$\phi_Y$	$\neg\phi_Y$
$\phi_X$	$-(v_{11})$	$--(v_{12})$
$\times$		
$\neg\phi_X$	$--(v_{21})$	$++(v_{22})$

The second type, denoted by sailor (with decomposition  $I_{\text{sailor}}$ ), is defined by the structural normal form of SEAMAN CLUMSY-like cases (with  $v_{11} < v_2 < v_{12} = v_{21}$ ), recall:

---

<sup>144</sup>As already mentioned (cf. Footnote 129, see also Figure 7.8), SEQUENTIAL CASES should automatically be treated in parallel because, against the background of the APPROACH, each COORDINATION CASE is ultimately a kind of superposition of several SEQUENTIAL CASES.

		Y	
		$\phi_Y$	$\neg\phi_Y$
		$\phi_X$	$\neg\phi_X$
sailor		$\phi_Y$	$\neg\phi_Y$
	$\phi_X$	---	( $v_{12}$ )
X		( $v_{21}$ )	- ( $v_{22}$ )

Annotated GEFs of trouble and sailor can be found in Figure 8.4.

A wide variety of criteria for good solutions were gathered throughout this book, which come together in this section (cf. especially Chapter 4 and see also Table 8.2). First, a truly objective-consequentialist theory in the sense of Definition 2.3 is sought, more precisely even a MOAC theory, which, of course, should be in particular consistent. Second, PMH in the form of a reasonably acceptable version (in the sense of Principle 8.1 in Section 8.1.2) is taken as a general guideline for the acceptance of suitable amendments, quite as explained in Section 8.2 in detail by considerations of policies. Third, the aim is to achieve NO MORAL DILEMMAS and, to a large extent, DEONTOIC COMPLETENESS, quite as suggested in Section 4.3. For it is necessary to close the deontic gaps diagnosed in Section 6.5 as causal for the REAL CHALLENGE (and thus to approach deontic completeness as MOAC) in a way that enables rightly acting, i.e., that does not introduce moral dilemmas, thus in particular also corresponds to RESOLVABILITY. Fourth, the touchstones mentioned here are intended to be a kind of *core test set* in the sense of (CORE) EXTENSIONAL ADEQUACY (cf. Principle 4.3), without introducing unnecessary types of entities nor unduly increasing complexity (in the sense of PARSIMONY and SIMPLICITY).

Three candidates remain to be considered: SUMMATION, MAXIMIZATION, and EXPECTED UTILITY. First, we start by comparing SUMMATION

<b>Requirements</b>	CONCEPTUAL DEONTIC CONSISTENCY, BEING A MOAC THEORY
<b>Desiderata</b>	RESOLVABILITY, NO MORAL DILEMMAS, STRONG NO MORAL DILEMMAS, WEAK DEONTIC COMPLETENESS, DEONTIC COMPLETENESS
<b>Comparative Cachets</b>	(CORE) EXTENSIONAL ADEQUACY, PARSIMONY, SIMPLICITY

Table 8.2: A non-final list of criteria as introduced in Chapter 4

and MAXIMIZATION. Let's call the policy that is induced by MOCoR together with SUMMATION  $\pi^\Sigma$  and the policy that is induced by MOCoR together with MAXIMIZATION  $\pi^{\max}$ .

First, we calculate the worth of  $\pi^\Sigma$ . For this case, I do this once in a very detailed manner, applying the various definitions and determinations from above, starting from Definition 8.4, and then plugin in the values from trouble. For the other combinations of text cases and amendments, I proceed in a less detailed way. So, let us start by recalling

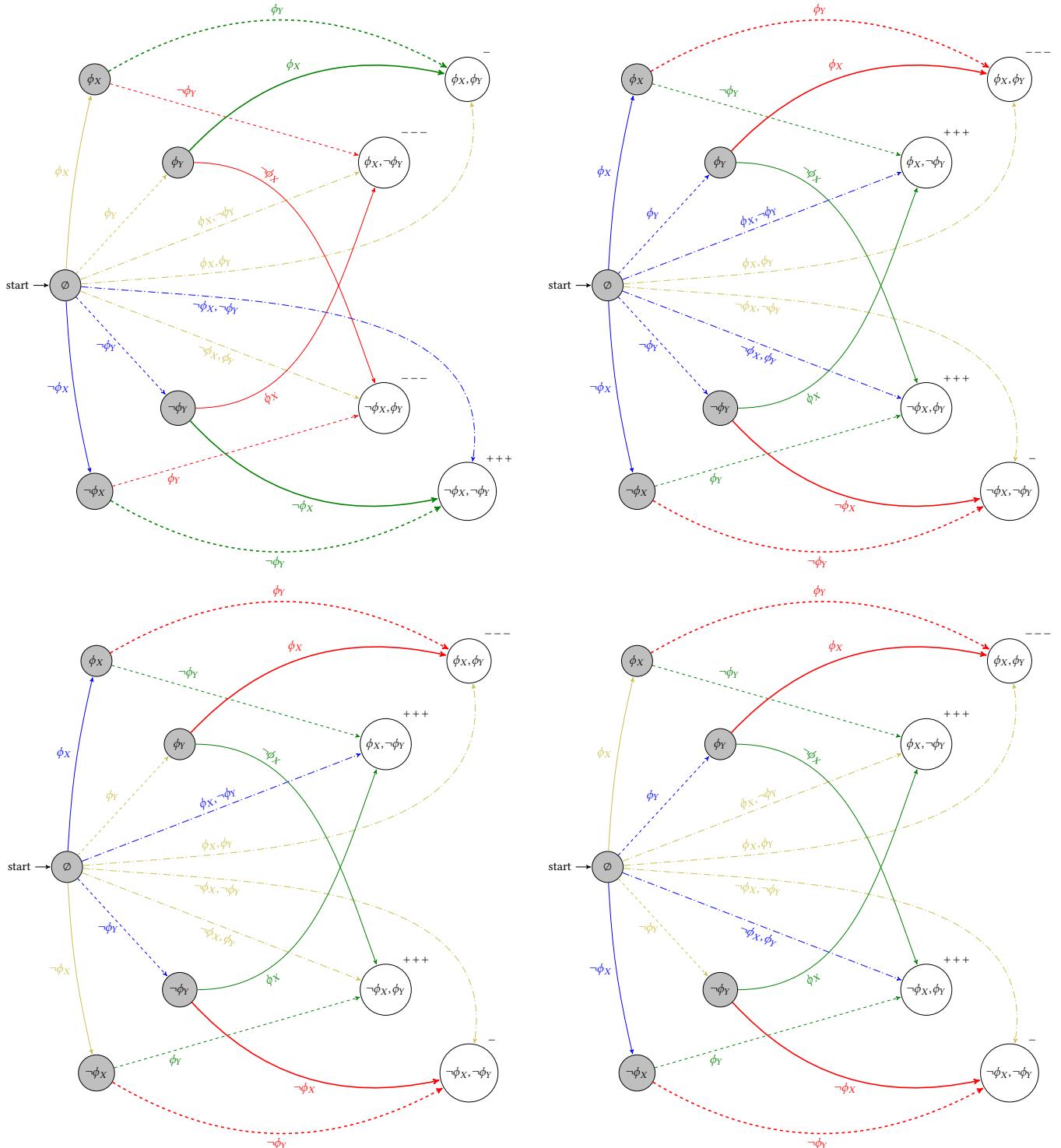
**Definition 7.2 (SUMMATION)** *Let  $D$  be a individual decision situation with  $D := \langle A, \Phi, \text{Out}_C : \Phi \rightarrow \mathcal{O} \rangle$  and actual context  $C$ . Further, let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function. The value of a individual decision situation  $D$  is*

$$\text{Val}^\Sigma(D) := \sum_{\phi \in \Phi} \text{Val}(\phi).$$

*Let  $D$  be a collective decision situation with actual context  $C$  and decomposition  $I_D := \{ D_{A_i} \mid A \in \mathcal{A}_D \}$  and let  $\text{Val} : \mathcal{W} \rightarrow \mathbb{R}$  be some valuation function.*

*The value of a collective decision situations to  $D$  is*

$$\text{Val}^\Sigma(D) := \sum_{D_{A_i} \in I_D} \text{Val}^\Sigma(D_{A_i}).$$

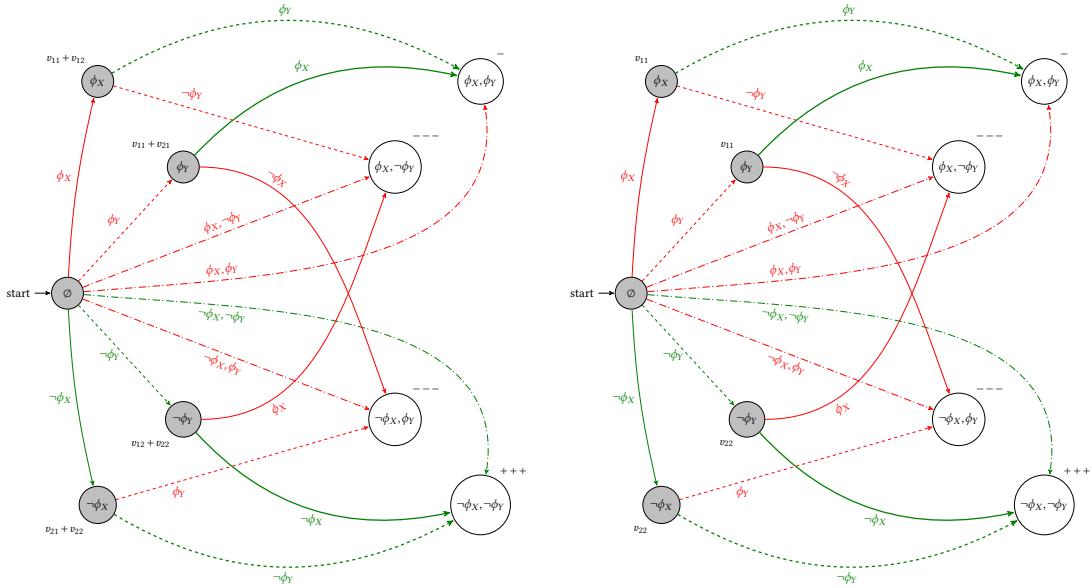


**Figure 8.4:** The GEFs trouble (top left) and for sailor (all other). Green and red edges show assessments MOCoR already provides us with, even without any amendments (green edges correspond to right and red edges to wrong actions in the sense of MOCoR). Blue edges correspond to edges we would *ideally* have marked as right by MOCoR together with an amendment, yellow ones such version of MOCoR would *ideally* have assessed as wrong. In the last chapter, we have already found evidence that all still considered amendments can deliver such ideal assessments for trouble-like cases. Even though it is not clear what actually would be ideal assignments for sailor-like cases, all three plausible candidates in light of PMH are shown here. However, we *cannot reasonably expect* to find an ideal solution for these cases, anyway: while the top-right version contradicts our principle of how moral assessments of individual options translate to the moral status of synchronous combinations of options, the other two would violate with NORMATIVE SUPERVENIENCE (cf. Section 8.1.2). The question remains: what is the best camp MOAC can reasonably hope to achieve?

We can now unfold:

$$\begin{aligned}
 \mathsf{Val}_{\text{trouble}}^{\Sigma}(\pi^{\Sigma}) &= \sum_{D \in I_{\text{trouble}}} \pi^{\Sigma}(D) \cdot \mathsf{Val}_D^{\Sigma}(\pi^{\Sigma}) \\
 &= \sum_{D \in I_{\text{trouble}}} \pi^{\Sigma}(D) \sum_{\phi \in \Phi_D} \pi^{\Sigma}(\phi) \cdot \mathsf{Val}_{D \downarrow \phi}^{\Sigma}(\pi^{\Sigma}) \\
 &= \sum_{D \in I_{\text{trouble}}} \pi^{\Sigma}(D) \sum_{\phi \in \Phi_D} \pi^{\Sigma}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \pi^{\Sigma}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi \downarrow \psi}^{\Sigma}(\pi^{\Sigma}) \\
 &= \sum_{D \in I_{\text{trouble}}} \pi^{\Sigma}(D) \sum_{\phi \in \Phi_D} \pi^{\Sigma}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \pi^{\Sigma}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi}^{\Sigma}(\psi) \\
 &= \sum_{D \in I_{\text{trouble}}} \pi^{\Sigma}(D) \sum_{\phi \in \Phi_D} \pi^{\Sigma}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \pi^{\Sigma}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi}(\psi) \\
 &= \sum_{D \in I_{\text{trouble}}} \frac{1}{2} \sum_{\phi \in \Phi_D} \Pr^{\Sigma}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \Pr^{\Sigma}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi}(\psi) \\
 &= \frac{1}{2} \left( \Pr^{\Sigma}(\phi_X) \left( \Pr^{\Sigma}(\phi_Y) \cdot \mathsf{Val}_{D \downarrow \phi_X}(\phi_Y) + \Pr^{\Sigma}(\neg\phi_Y) \cdot \mathsf{Val}_{D \downarrow \phi_X}(\neg\phi_Y) \right) \right. \\
 &\quad \left. + \Pr^{\Sigma}(\neg\phi_X) \left( \Pr^{\Sigma}(\phi_Y) \cdot \mathsf{Val}_{D \downarrow \neg\phi_X}(\phi_Y) + \Pr^{\Sigma}(\neg\phi_Y) \cdot \mathsf{Val}_{D \downarrow \neg\phi_X}(\neg\phi_Y) \right) \right) \\
 &+ \frac{1}{2} \left( \Pr^{\Sigma}(\phi_Y) \left( \Pr^{\Sigma}(\phi_X) \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\phi_X) + \Pr^{\Sigma}(\neg\phi_X) \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\neg\phi_X) \right) \right. \\
 &\quad \left. + \Pr^{\Sigma}(\neg\phi_Y) \left( \Pr^{\Sigma}(\phi_X) \cdot \mathsf{Val}_{D \downarrow \neg\phi_Y}(\phi_X) + \Pr^{\Sigma}(\neg\phi_X) \cdot \mathsf{Val}_{D \downarrow \neg\phi_Y}(\neg\phi_X) \right) \right) \\
 &= \frac{1}{2} \left( 0 \cdot \left( 1 \cdot \mathsf{Val}_{D \downarrow \phi_X}(\phi_Y) + 0 \cdot \mathsf{Val}_{D \downarrow \phi_X}(\neg\phi_Y) \right) \right. \\
 &\quad \left. + 1 \cdot \left( 0 \cdot \mathsf{Val}_{D \downarrow \neg\phi_X}(\phi_Y) + 1 \cdot \mathsf{Val}_{D \downarrow \neg\phi_X}(\neg\phi_Y) \right) \right) \\
 &+ \frac{1}{2} \left( 0 \cdot \left( 1 \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\phi_X) + 0 \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\neg\phi_X) \right) \right. \\
 &\quad \left. + 1 \cdot \left( 0 \cdot \mathsf{Val}_{D \downarrow \neg\phi_Y}(\phi_X) + 1 \cdot \mathsf{Val}_{D \downarrow \neg\phi_Y}(\neg\phi_X) \right) \right) \\
 &= \frac{1}{2} \mathsf{Val}_{D \downarrow \neg\phi_X}(\neg\phi_Y) + \frac{1}{2} \mathsf{Val}_{D \downarrow \neg\phi_Y}(\neg\phi_X) \\
 &= \frac{1}{2} v_{22} + \frac{1}{2} v_{22} = v_{22}
 \end{aligned}$$

All this is just a precise, formal, and particular general way to show how following  $\pi^{\Sigma}$  through a trouble-like case leads to the best outcome. Slightly more precise, there are two ways how things can unfold, both having a 50% chance to occur. Each agent could act first and, following MOCOR plus SUMMATION, would perform their option  $\neg\phi$ . In other words, they would choose the second row and column, respectively, because the sum of  $v_{11}$  and  $v_{12}$  is less than the sum of  $v_{21}$  and  $v_{22}$  (and the sum of  $v_{11}$  and  $v_{21}$  is less than



**Figure 8.5:** The annotated GEFs for trouble for MOCor with SUMMATION (left) and for MOCor with MAXIMIZATION. Both solutions are ideal (cf. Figure 8.4).

the sum of  $v_{12}$  and  $v_{22}$ ). The value of the policy  $\pi_\Sigma$  for each of the resulting individual decision situations, thus, equals two times the half of the value of the best possible outcome, namely  $v_{22}$ .

Now compare this to the value of  $\pi_{\max}$ :

$$\begin{aligned}\text{Val}_{\text{trouble}}^{\max}(\pi^{\max}) &= \sum_{D \in I_{\text{trouble}}} \pi^{\max}(D) \cdot \text{Val}_D^{\max}(\pi^{\max}) \\ &= \frac{1}{2} \text{Val}_{D_{\neg\phi_X}}(\neg\phi_Y) + \frac{1}{2} \text{Val}_{D_{\neg\phi_Y}}(\neg\phi_X) \\ &= \frac{1}{2} v_{22} + \frac{1}{2} v_{22} = v_{22}\end{aligned}$$

Accordingly, both amendments give the same result for traditional TROUBLEMAKERS and, thus, yield paths in the GEF of trouble (cf. Figure 8.5). The results are ideal and thus do not only fill the normative gaps, hence they help MOAC to master the REAL CHALLENGE.

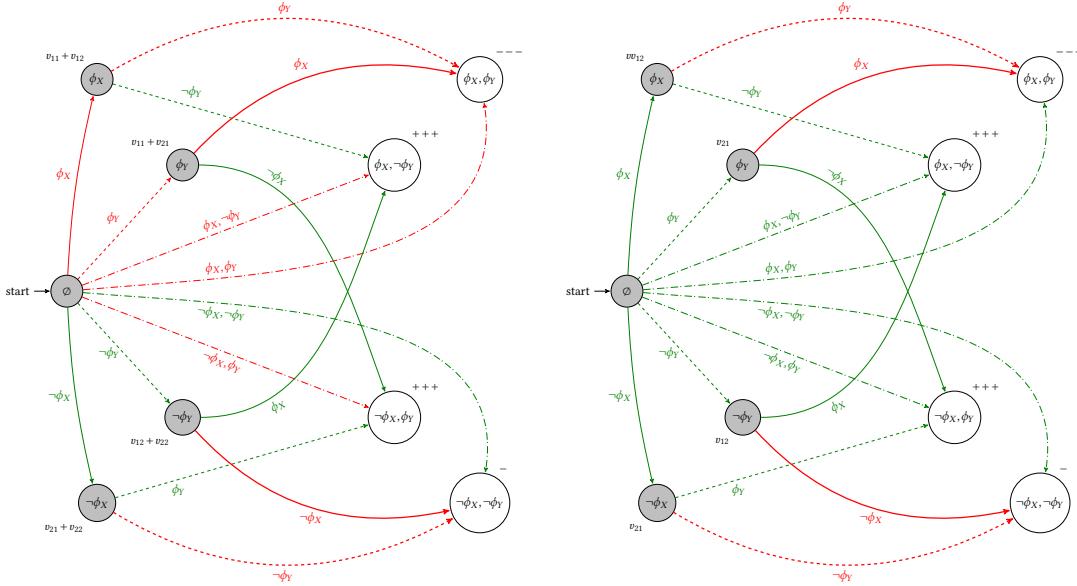
However, the two amendments are interestingly different for sailor. First, let us observe that both policies have, again, the same (expected) value:

$$\begin{aligned}
 \mathsf{Val}_{\text{sailor}}^{\Sigma}(\pi^{\Sigma}) &= \sum_{D \in I_{\text{sailor}}} \pi^{\Sigma}(D) \cdot \mathsf{Val}_D^{\Sigma}(\pi^{\Sigma}) \\
 &= \sum_{D \in I_{\text{sailor}}} \frac{1}{2} \sum_{\phi \in \Phi_D} \Pr^{\Sigma}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \Pr^{\Sigma}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi}(\psi) \\
 &= \frac{1}{2} \mathsf{Val}_{D \downarrow \neg \phi_X}(\neg \phi_Y) + \frac{1}{2} \mathsf{Val}_{D \downarrow \neg \phi_Y}(\neg \phi_X) \\
 &= \frac{1}{2} v_{22} + \frac{1}{2} v_{22} = v_{22}
 \end{aligned}$$

Next, we compute the result for  $\pi^{\max}$ :

$$\begin{aligned}
 \mathsf{Val}_{\text{sailor}}^{\max}(\pi^{\max}) &= \sum_{D \in I_{\text{trouble}}} \pi^{\max}(D) \cdot \mathsf{Val}_D^{\max}(\pi^{\max}) \\
 &= \sum_{D \in I_{\text{sailor}}} \frac{1}{2} \sum_{\phi \in \Phi_D} \Pr^{\max}(\phi) \sum_{\psi \in \Phi_{D \downarrow \phi}} \Pr^{\max}(\psi) \cdot \mathsf{Val}_{D \downarrow \phi}(\psi) \\
 &= \frac{1}{2} \left( \Pr^{\max}(\phi_X) \left( \Pr^{\max}(\phi_Y) \cdot \mathsf{Val}_{D \downarrow \phi_X}(\phi_Y) + \Pr^{\max}(\neg \phi_Y) \cdot \mathsf{Val}_{D \downarrow \phi_X}(\neg \phi_Y) \right) \right. \\
 &\quad \left. + \Pr^{\max}(\neg \phi_X) \left( \Pr^{\max}(\phi_Y) \cdot \mathsf{Val}_{D \downarrow \neg \phi_X}(\phi_Y) + \Pr^{\max}(\neg \phi_Y) \cdot \mathsf{Val}_{D \downarrow \neg \phi_X}(\neg \phi_Y) \right) \right) \\
 &\quad + \frac{1}{2} \left( \Pr^{\max}(\phi_Y) \left( \Pr^{\max}(\phi_X) \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\phi_X) + \Pr^{\max}(\neg \phi_X) \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\neg \phi_X) \right) \right. \\
 &\quad \left. + \Pr^{\max}(\neg \phi_Y) \left( \Pr^{\max}(\phi_X) \cdot \mathsf{Val}_{D \downarrow \neg \phi_Y}(\phi_X) + \Pr^{\max}(\neg \phi_X) \cdot \mathsf{Val}_{D \downarrow \neg \phi_Y}(\neg \phi_X) \right) \right) \\
 &= \frac{1}{2} \left( \frac{1}{2} \cdot \left( 0 \cdot \mathsf{Val}_{D \downarrow \phi_X}(\phi_Y) + 1 \cdot \mathsf{Val}_{D \downarrow \phi_X}(\neg \phi_Y) \right) \right. \\
 &\quad \left. + \frac{1}{2} \cdot \left( 1 \cdot \mathsf{Val}_{D \downarrow \neg \phi_X}(\phi_Y) + 0 \cdot \mathsf{Val}_{D \downarrow \neg \phi_X}(\neg \phi_Y) \right) \right) \\
 &\quad + \frac{1}{2} \left( \frac{1}{2} \cdot \left( 0 \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\phi_X) + 1 \cdot \mathsf{Val}_{D \downarrow \phi_Y}(\neg \phi_X) \right) \right. \\
 &\quad \left. + \frac{1}{2} \cdot \left( 1 \cdot \mathsf{Val}_{D \downarrow \neg \phi_Y}(\phi_X) + 0 \cdot \mathsf{Val}_{D \downarrow \neg \phi_Y}(\neg \phi_X) \right) \right) \\
 &= \frac{1}{4} \mathsf{Val}_{D \downarrow \phi_X}(\neg \phi_Y) + \frac{1}{4} \mathsf{Val}_{D \downarrow \neg \phi_X}(\phi_Y) + \frac{1}{4} \mathsf{Val}_{D \downarrow \phi_Y}(\neg \phi_X) + \frac{1}{4} \mathsf{Val}_{D \downarrow \neg \phi_Y}(\phi_X) \\
 &= \frac{1}{4} v_{22} + \frac{1}{4} v_{22} + \frac{1}{4} v_{22} + \frac{1}{4} v_{22} = v_{22}
 \end{aligned}$$

Although the (expected) values of the two policies `sailor`-like cases, a closer look reveals that they are importantly different with respect to the paths they allow for. Interestingly, both are unsatisfactory in that `SUMMATION` is too restrictive, and `MAXIMIZATION` is too permissible (cf. Figure 8.6 helps immensely understand the point I make here). This is because, according to  $\pi^{\Sigma}$ , there are combinations of simultaneous actions that contain wrong actions



**Figure 8.6:** The annotated GEFs for sailor for MOCoR with SUMMATION (left) and for MOCoR with MAXIMIZATION. Both solutions are *not* ideal (cf. Figure 8.4).

(i.e., red paths in the GEF) that lead to optimal results. As a consequence, it is possible that, according to the resulting MOAC theory, even though the best possible outcome is achieved, one agent did wrong (imagine, for instance, Mikel gets the buoy at the very same moment Laika jumps, or vice versa). Thus, it can happen that the optimal outcome is achieved, but one agent (namely X) has acted wrongly. According to  $\pi^{\max}$ , on the other hand, all first actions are right, which has the consequence, among other things, that there are several combinations of simultaneous actions that are right but lead to suboptimal outcomes (imaging Maika and Laika jump – or get the buoy – in the very same moment).<sup>145</sup> While I do not consider these objections to be conclusive because they are based on the borderline case of simultaneous action, they nevertheless speak *pro tanto* against the two amendments.

That leaves us with MIXED STRATEGIES as a candidate amendment. In the following, then, the program outlined by J.C.C. Smart in response

<sup>145</sup>A lexicographic version of  $\pi^{\max}$  (cf. Footnote 135) would yield the same results as  $\pi^\Sigma$ .

to Brandt must be executed (cf. Section 7.3.2.3). We start by explicating the function describing the expected value as a function of the chosen probability distribution.<sup>146</sup> This is for both types of testbed cases with the same general formula, where  $p$  is the probability for performing the corresponding action  $\phi$  (exploiting that  $v_{12} = v_{21}$  in both types of testbed cases):

$$\begin{aligned} EV(p) &= p^2v_{11} + p(1-p)v_{12} + p(1-p)v_{21} + (1-p)^2v_{22} \\ &= p^2v_{11} + 2p(1-p)v_{12} + (1-p)^2v_{22} \\ &= p^2v_{11} + 2pv_{12} - 2p^2v_{12} + (1-2p+p^2)v_{22} \\ &= p^2v_{11} + 2pv_{12} - 2p^2v_{12} + v_{22} - 2pv_{22} + p^2v_{22} \\ &= p^2(v_{11} - 2v_{12} + v_{22}) + 2p(v_{12} - v_{22}) + v_{22} \end{aligned}$$

We are interested in the maximal value for  $EV$ . We thus consider the first derivative:

$$EV'(p) = 2p(v_{11} - 2v_{12} + v_{22}) + 2(v_{12} - v_{22})$$

Next, we set this first derivative equal to zero and solve for  $p$ :

$$\begin{aligned} EV'(p) &= 0 \\ \Leftrightarrow & 0 = 2p(v_{11} - 2v_{12} + v_{22}) + 2(v_{12} - v_{22}) \\ \Leftrightarrow & 0 = p(v_{11} - 2v_{12} + v_{22}) + v_{12} - v_{22} \\ \Leftrightarrow & v_{22} - v_{12} = p(v_{11} - 2v_{12} + v_{22}) \\ \text{thus: } & p_{\text{extrem}} = \frac{\overbrace{v_{22} - v_{12}}^{=:x}}{\underbrace{v_{11} - 2v_{12} + v_{22}}_{=:y}} \end{aligned}$$

---

<sup>146</sup>Note that in light of NORMATIVE SUPERVENIENCE and the fact that both types of cases in the testbed are symmetric, this distribution must be equal for both agents.

We compute the two boundary values and calculate the extreme value:

$$EV(0) = v_{22}$$

$$EV(1) = v_{11}$$

$$\begin{aligned} EV(p_{\text{extrem}}) &= \left(\frac{x}{y}\right)^2 \underbrace{(v_{11} - 2v_{12} + v_{22})}_{=y} + 2\frac{x}{y} \underbrace{(v_{12} - v_{22})}_{=-x} + v_{22} \\ &= \frac{x^2}{y} - 2\frac{x^2}{y} + v_{22} \\ &= v_{22} - \frac{x^2}{y} \\ &= v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} \end{aligned}$$

Since  $v_{11} < v_{22}$  for both testbad cases and since  $v_{22}$  equals the value of the two testbad cases for both **SUMMATION** and **MAXIMIZATION**, we can already conclude that **MIXED STRATEGIES** can at least match the value of these two amendments (because  $EV(0) = v_{22}$ ). What we want to know at this point is, thus, whether the extreme value  $EV(p_{\text{extrem}})$  is even higher than  $v_{22}$ . With this general solution at hand, we can thus now turn to the two testbed cases one after another.

We start with **trouble**, which is characterized by the following order of the values of the final outcomes:

$$v_{12} = v_{21} < v_{11} < v_{22}$$

For **trouble**, however, we can skip at this point, as we know that  $v_{22}$  is actually the ideal solution. As a result, MOCOR together with **MIXED STRATEGIES** masters classical **TROUBLEMAKERS** (as did MOCOR together with **SUMMATION** and **MAXIMIZATION**, respectively). Thus, **MIXED STRATEGIES** solves not only the **REAL CHALLENGE** but also the **TROUBLEMAKER-based CHALLENGE**.

So we turn to **sailor**, which is characterized by the following order of the values of the final outcomes:

$$v_{11} < v_{22} < v_{12} = v_{21}$$

We can rewrite this to

$$v_{11} - v_{12} < v_{22} - v_{12} < 0$$

We check for

$$\begin{aligned} & v_{22} - \frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} > ? & v_{22} \\ \Leftrightarrow & -\frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}} > 0 \\ \Leftrightarrow & \overbrace{\frac{(v_{22} - v_{12})^2}{v_{11} - 2v_{12} + v_{22}}}^{>0} < 0 \\ \Leftrightarrow & v_{11} - 2v_{12} + v_{22} < 0 \\ \Leftrightarrow & v_{11} - v_{12} + v_{22} - v_{12} < 0 \end{aligned}$$

Thanks to the above rewriting of the case defining order, we know that

$$v_{11} - v_{12} < 0$$

and that

$$v_{22} - v_{12} < 0.$$

So we also know that

$$\overbrace{(v_{11} - v_{12})}^{<0} + \overbrace{(v_{22} - v_{12})}^{<0} < 0$$

Thus, **MIXED STRATEGIES** performs even better than **SUMMATION** and **MAXIMIZATION** for **SEAMAN CLUMSY**-like cases.

The results of all the computations in this section can be found in Table 8.3, and they allow me to come to a final judgment.

### 8.3. THE FINAL EVALUATION

365

Amendment	Best for all cases	Guarantees best for classical TROUBLEMAKERS	Guarantees best for classical SEAMAN CLUMSY-like cases	Could lead to best in SEAMAN CLUMSY-like cases	Guaranteed not worst in SEAMAN CLUMSY-like cases	Best EV for SEAMAN CLUMSY-like cases	further pros	further cons
SUMMATION	✗	✓	✗	✗	✓	✗	simple	assesses some optimal paths as wrong, thus assesses <i>combinations</i> rather too restrictively
MAXIMIZATION	✗	✓	✗	✓	✗	✗	connects well with the »MOAC«, all optimal paths are assessed as right	assesses <i>some combinations</i> too permissibly
MAXIMIZATION [lexicographic]	✗	✓	✗	✗	✓	✗	connects well with the »MOAC«	assesses some optimal paths as wrong, thus assesses <i>combinations</i> rather too restrictively
MIXED STRATEGIES	✓	✓	✗	✓	✗	✓	based on a well-established concept from instrumental rationality	required adding a new formal methodology to the consequentialist toolbox

Table 8.3: All three candidate amendments cope perfectly with TROUBLEMAKERS. MIXED STRATEGIES copes best with SEAMAN CLUMSY-like cases, but also comes with the price that the catastrophic result may occur. SUMMATION and lexicographic MAXIMIZATION lead to qualitatively identical outcomes.

### 8.3.2 And the Winner Is ...

The results are *not* conclusive in terms of what is the overall best amendment. Classical TROUBLEMAKERS can be handled ideally by all three amendments, while SEAMAN CLUMSY-like cases cannot be solved ideally by any of the three amendments. This, however, we cannot demand in light of the results in Section 8.1. Therefore, the final choice ultimately depends on how risk-averse or how risk-affine one is: Where MIXED STRATEGIES allows to achieve the optimum in SEAMAN CLUMSY, it also runs the risk of leading into the word case scenario.

An example with concrete values is helpful for illustration. Let's assume that the death of all three sailors would result in a disvalue of  $-3000$ . Conversely, a rescue of Clumsy would have a value of  $3000$ . A drowning of Clumsy and one further crew member would have a disvalue of  $-1000$ . This implies the following instance of SEAMAN CLUMSY:

		Laika	
		jump	get buoy
		jump	get buoy
Mike	jump	-3000	+3000
	get buoy	+3000	-1000

Naïve MAXIMIZATION allows for either action. So, given that the agents choose by fair randomization between the right actions, this would mean a 25% chance of every outcome, resulting in an expected value of  $-750 + 750 + 750 - 250 = 500$ . Lexicographic MAXIMIZATION and SUMMATION both result in the rightness of getting the buoy for both agents and, thus, in a

certain disvalue of -1000. MIXED STRATEGIES results in a probability of

$$p := \frac{-1000 - 3000}{-3000 - 2 \cdot 3000 - 1000} = \frac{-4000}{-10000} = \frac{2}{5}$$

and thus<sup>147</sup> an expected value of

$$\begin{aligned} EV(p) &:= -\left(\frac{2}{5}\right)^2 \cdot 3000 + 2\frac{2}{5} \cdot \frac{3}{5} \cdot 3000 - \left(\frac{3}{5}\right)^2 \cdot 1000 \\ &= -\frac{4}{25} \cdot 3000 + \frac{12}{25} \cdot 3000 - \frac{9}{25} \cdot 1000 \\ &= -480 + 1440 - 360 \\ &= 600. \end{aligned}$$

Interestingly, a naïve MAXIMIZATION approach actually provides a better (expected) value than a lexicographic approach in this case (simply because two comparatively very good alternative results remain possible with it – in general, of course, this observation does not apply). MIXED STRATEGIES, however, provides the highest value. At the same time, there is a 16% chance to result in the worst outcome, where all jump and die, as well as a 36% chance to result in the second-worst outcome, where Mikel and Laika both get the buoy and Clumsy drown in the meantime.

So, what is the final verdict? Ultimately, I believe that the usual arguments for MIXED STRATEGIES can be put forward that otherwise support expected utility calculations in the context of rational decision-making under risk. In particular, when applied to the totality of all (or even ›just‹ to

---

<sup>147</sup>Note that this aligns with the general formula derived and used above:

$$EV(p) = \underbrace{-1000}_{=v_{22}} - \underbrace{\frac{(-1000 - 3000)^2}{-1000 - 2 \cdot 3000 - 3000}}_{=v_{11} - 2v_{12} + v_{22}} = -1000 + \frac{16 \cdot 1000^2}{10000} = 600$$

the totality of the actually occurring) situations, MIXED STRATEGIES will pay off morally in aggregate. And the aggregate of the moral good is what matters for camp MOAC most. I thus advocated for MIXED STRATEGIES as the collective amendment of choice for MOAC. In my view, this even justifies modifying MOCOR to make room for maximizing expected value (cf. Section 7.3.2.3) and, accordingly, I suggest replacing MOCOR by

**Principle 7.1 (MOCOR (EU))** *Let  $D$  be a collective decision situation of agents  $\mathcal{A}$  with domain  $\Psi$  and actual context  $C$ .*

$\phi \in \Phi_A$  is right for agent  $A \in \mathcal{A}$  (given  $C$ ) if and only if  $A$ 's choice followed a probability distribution  $\text{Pr}_{\Phi_A} : \Phi_A \rightarrow [0, 1]$  where  $\text{Pr}_{\Phi_A}$  belongs to a set of probability distributions of all agents such that  $\text{Pr}_\Psi : \Psi \rightarrow [0, 1]$  defined as

$$\text{Pr}_\Psi(\Upsilon) := \prod_{A \in \mathcal{A}_\Upsilon} \text{Pr}_{\Phi_A}(\Upsilon_A)$$

maximize  $EV(\text{Pr}_\Psi)$ , relative to  $C$ .

I shall call the families of theories subscribing to this criterion of rightness MULTI-AGENT CONSEQUENTIALISM theories. MAC theories are fit for multi-agent scenarios, and it must be noted that for ›normal‹ decision situations (i.e., individual decision situations where actions immediately result in final outcomes), MOCOR and MOCOR (EU) are extensionally equivalent. Therefore, first, for ›traditional‹ individual decision situations, this theoretical adjustment cannot lead to new challenges; and second, MAC should be understood as a *generalization* of MOAC. In particular, MAC (and thus MOAC) remains *extensionally different* from decision-theoretic MSAC.

Of course, MIXED STRATEGIES comes with some conceptual challenges. On the one hand, one has to accept that not only the explicitly available actions exist but that it is at least in principle possible for agents to act according

to arbitrary probability distributions. On the other hand, one has to accept that acting according to such probability distributions now becomes the default of doing the right thing. However, I don't think that should be taken as a specific challenge to my approach. In fact, EU-MOAC is in the best of company. For instance, in instrumental rationality, it is commonplace to accept mixed strategies, and dismissing probabilistic approaches to the selection of actions would result in *massive* losses in expected value and, thus, to losses in the performance of our best theories of instrumental rationality. Therefore, I see no *genuine* reason against **MIXED STRATEGIES** as an amendment of choice for camp MOAC. I, therefore, argue for the acceptance of **MIXED STRATEGIES** as an amendment in the sense of reasonable MH to solve the **REAL CHALLENGE** without getting into a justified variant of **CHALLENGE**.

Before I retrace the success of my approach in the context of an overall conclusion, one comment concerning the non-exhaustiveness of my investigation: Of course, it cannot be excluded that the consideration of further amendments against the background of other, possibly more extensive and more complex collective decision situations turn out to be advantageous. Further, it cannot be excluded that different amendments for different (classes of) collective decision situations turn out to be advantageous. In this case, camp MOAC should consider an *ensemble solution*, i.e., incorporate several amendments that are used or combined depending on the structure of the specific collective decision situation to be assessed. Ensemble solutions have already shown promise in formal decision theory, especially in the area of machine learning (cf. Dong et al. 2020). A family of theories that has been unafraid of genuine conditional assessment for so long (we remember the CSM) is unlikely to shy away from the conditional use of amendments.

## 8.4 On Overall Success

In the light of Chapter 6 the REAL CHALLENGE had to be solved first before we could even turn back to the PYRAMID and thus to the CHALLENGE in all its variety. We remember the related argument:

### **Argument:** The REAL ARGUMENT

$P_{\exists NS}$ : There are non-separable collective decision situations, i.e., collective decision situations in which, relative to the actual and normatively sufficiently complete context, it is not defined for at least one agent with respect to at least two of their options how they are to be ranked with respect to the moral quality of their actual consequences.

$P_{\exists GAPS}$ : If there are such collective action situations, then there is a widespread class of decisions to which MOAC has nothing illuminating to contribute; that is, there are a myriad of systematic deontic gaps.

$P_{\text{No GAPS}!}$ : If a moral theory is adequate, then there are no widespread class of decisions to which the theory has nothing illuminating to contribute; that is, there are no systematic deontic gaps.

---

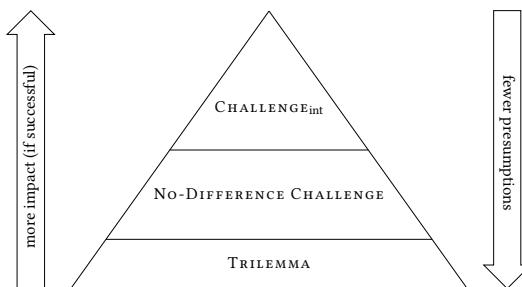
$C_{\neg \text{Adeq}}$ : MOAC is not adequate.

Thanks to the APPROACH,  $P_{\exists NS}$  has long been off the table, and COMPOSITIONALISM has proven to be defensible. Each action of each agent has a properly defined consequence, namely the decision situation that remains

for all the other agents, given that every action of this very agent. However, this only helps to a limited extent. For as long as we cannot rank the newly discovered consequences, the existence of a multitude of actions that are not rankable with respect to the moral qualities of their outcomes remains, and so normative gaps prevail. Only by adding a collective amendment is it possible for camp MOAC to master the REAL CHALLENGE.

This raised the question of how the gaps *should* be closed, i.e., the question of the specific choice of 'the right' collective amendment. Given what we learned about the CHALLENGE in the first part of this project, it was natural to choose an amendment that would not immediately raise the CHALLENGE again. Looking at the PYRAMID, we concentrated our considerations for the time being on the CHALLENGE as CHALLENGE<sub>int</sub> and thus PMH. We recall

**Argument:** The ARGUMENT (tentative)



**Figure 8.7:** The three variants that make up the CHALLENGE as the PYRAMID as tackled in this thesis, in order of strength and number of preconditions they require.

*P<sub>3T</sub>:* There are TROUBLEMAKERS: collective decision situations in which the agents can act in ways such that together they would produce a morally suboptimal outcome, but none of them could make a difference for the better by unilaterally acting differently.

$P_{MOC_{\text{O}}R}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{MH}$ : If a moral theory is adequate, then, if the agents in a collective decision situation were to act in ways such that together they would not produce a morally optimal outcome, then (necessarily) at least one of the agents acted wrongly (i.e., according to this theory).

---

$C_{\neg \text{Adeq}}$ : MOAC is not an adequate moral theory.

Given a collective amendment and the consequences revealed by the APPROACH, there is a plausible reading of  $P_{MOC_{\text{O}}R}$  that does not involve the strange ex-post reasoning that originally undermined the ARGUMENT's validity. The task, then, was to find an amendment that would allow camp MOAC to condemn some action in every combination that leads to suboptimal results and, thus, an amendment that does not violate the condition on the rightmost side of  $P_{MH}$ .

But before we actually accept MH (and thus  $P_{MH}$ ) and direct all our efforts to it, it was necessary to think about PMH one more time. In fact, we found with SEAMAN CLUMSY in Section 8.1 a non-trivial counterexample to MH: Astoundingly, we found that a theory satisfying MH *cannot* be an objectively consequentialist theory because, in SEAMAN CLUMSY-like cases, such a theory would necessarily violate the principle NORMATIVE SUPERVENIENCE implied by the very definition of objective consequentialism.

In the end, we thus introduced a restriction of MH to the *reasonably expectable best results*, a restriction that we succeeded in substantiating by making fruitful the notion of *theory-induced policies* and their *expected value relative to (sets of) decision situations*. In the end, there was a winner, MIXED STRATEGIES as the collective amendment to be chosen in terms of MOAC.

In this respect, the CHALLENGE may also be regarded as being mastered as the CHALLENGE<sub>int</sub>. We remember the corresponding solution space (cf. Figure 8.8). Given the amendment of choice (cf. Section 8.2.2), MIXED STRATEGIES, yields a modified version of MO- CoR called MOCOR (EU), ultimately defining MULTI- AGENT CONSEQUENTIALISM (shorter: MAC). In traditional, >Hi-Lo< TROUBLE- MAKERS such as TWO FACTERIES or WHIFF AND POOF, MAC necessarily identifies wrong actions in combinations that result in sub-optimal outcomes and even

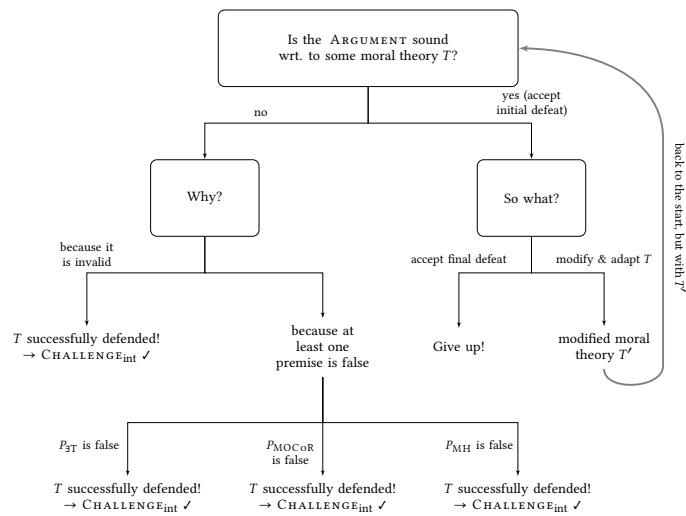


Figure 8.8: The solution space of the CHALLENGE as CHALLENGE<sub>int</sub>.

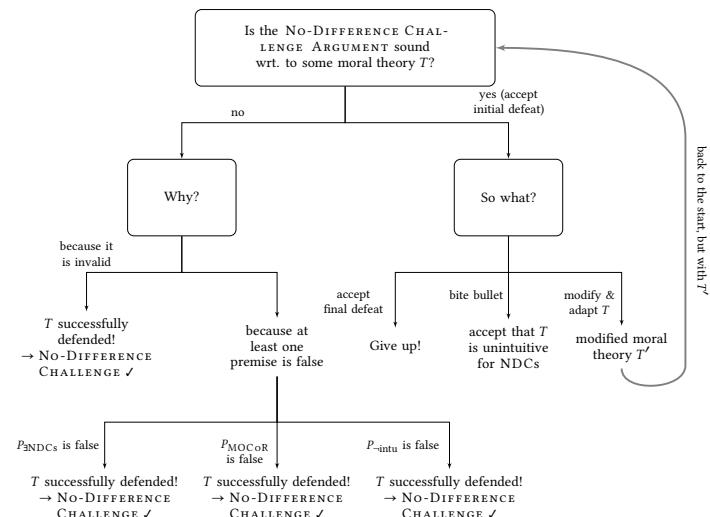


Figure 8.9: The solution space of the CHALLENGE as No-DIFFERENCE CHALLENGE.

recommends only actions that, if followed, necessarily lead to one of the best possible outcomes.

For more demanding cases like SEAMAN CLUMSY with more than one optimum, which furthermore does not lie on the main diagonal, MAC may even lead to the worst possible result, but that would then be bad moral luck. A price worth paying given that this choice of amendment arguably *maximizes the total expected value*, i.e., in the long run, that is, over all (possible or actual) decision situations (of the investigated types, at least). Referring to the solution space of CHALLENGE<sub>int</sub>, MAC (strictly speaking) takes two exits at once: both  $P_{MOC_{oR}}$  and  $P_{MH}$  are false – both times because the »necessarily« is not warranted. However, if restricted to TROUBLEMAKERS, MAC just ›falsifies‹  $P_{MOC_{oR}}$ .

This exit helps a lot when we arrive at the second level of the PYRAMID, the CHALLENGE as NO-DIFFERENCE CHALLENGE. We recall the associated solution space (cf. Figure 8.9) and the corresponding argument:

**Argument:** The NO-DIFFERENCE CHALLENGE ARGUMENT (tentative)

$P_{\exists NDC_s}$ : There are NO-DIFFERENCE CASES: collective decision situations in which there is at least one agent that can act in ways such that it seems intuitively morally wrong, but the agent could not make a difference for the morally better by unilaterally acting differently.

$P_{MOC_{oR}}$ : If the agents in a collective decision situation were to act in ways such that none of them could make a difference for the better by unilaterally acting differently, then (necessarily) each of them would act morally right (according to MOAC).

$P_{\neg \text{intu}}$ : If a moral theory assesses intuitively morally wrong actions as morally right for a significant class of situations, then  $T$  is counterintuitive.

---

$C_{\neg \text{intu}}$ : MOAC is counterintuitive.

Again,  $P_{\text{MOCOr}}$  is false according to MAC, and so the NO-DIFFERENCE CHALLENGE is solved automatically.

However, MAC might violate some other intuitions, maybe even MOAC-inspired intuitions. For instance, the recommendation to act in a certain way in cases like SEAMAN CLUMSY with specific probabilities could possibly be seen as non-intuitive for other reasons and I have indeed sometimes been met with incredulous astonishment when I have presented my approach in recent years for this very reason). I think, however, that ›informed and equalized‹ intuitions should have no problem with this, at least ›informed and equalized‹ consequentialist intuition, and thus camp MOAC should just follow the way of formal decision theory and simply accept this possibility and the power that comes with it. After all, even consequentialists would also have to choose *somewhat* between several right options – even if this case of two exactly morally equivalent consequences may sound exotic to consequentialist ears. If one insists on not swallowing this probabilistic frog, one can still turn to SUMMATION or the naïve or lexicographic variant of MAXIMIZATION. Then, one would have to accept the loss of some expected value in SEAMAN CLUMSY-like cases but get around the acceptance of apparently brain-twisting probabilistic actions. I don't see any reason for that – but either way, it may make the choice facing the champions of camp MOAC in NO-DIFFERENCE CASES much more pleasant for some of these champions.

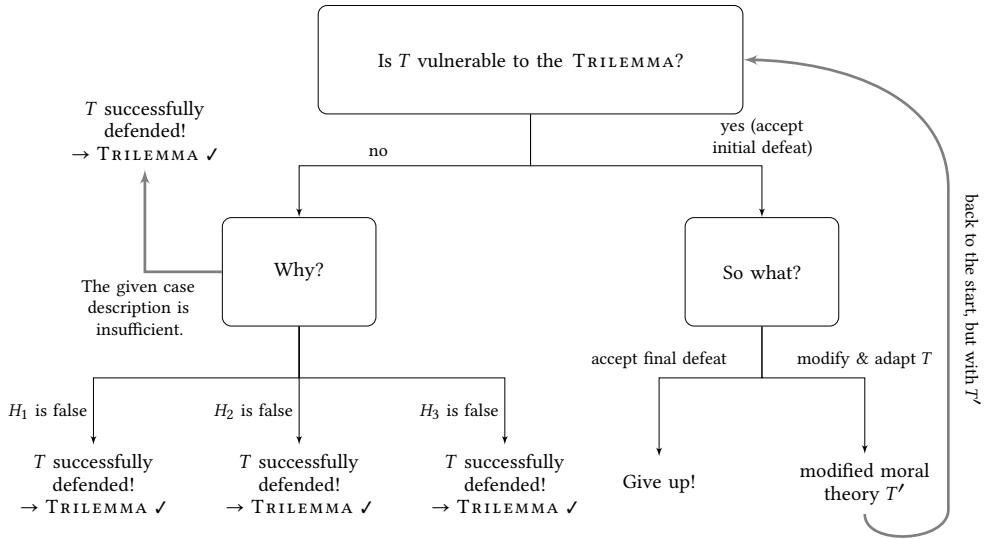


Figure 8.10: The solution space of the TRILEMMA.

In the end, this only leaves us with the TRILEMMA. We recall its solution space (cf. Figure 4.4). To solve the TRILEMMA required to reject one of the following three propositions :

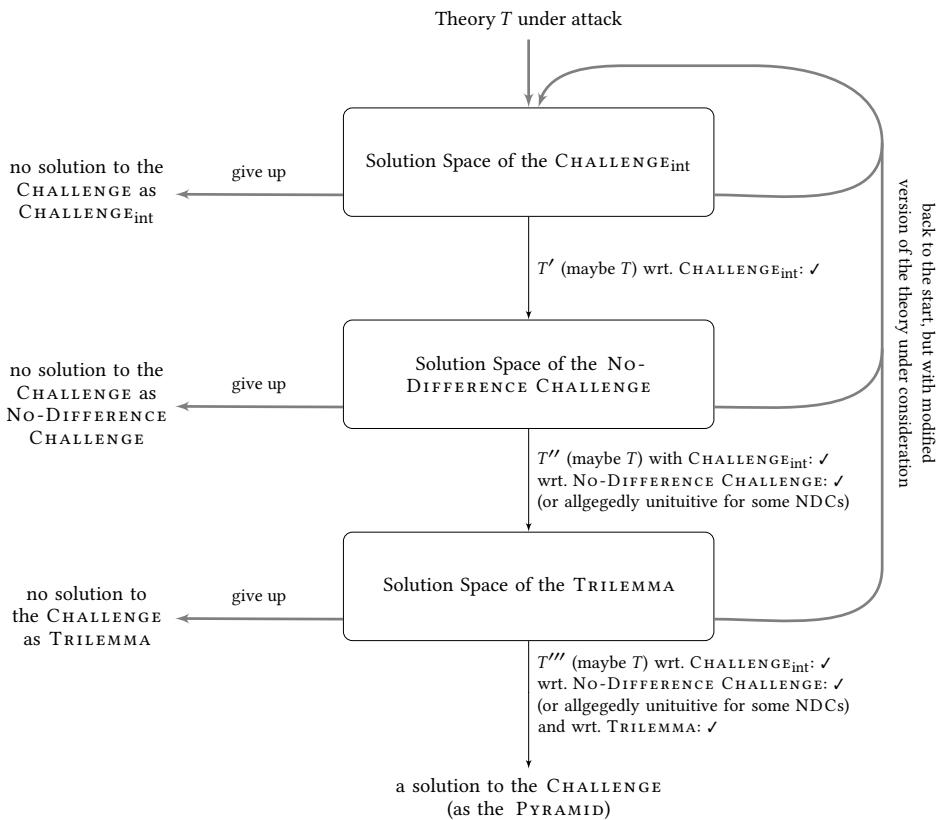
( $H_1$ ) What Ann and Ben have done is morally wrong.

( $H_2$ ) If something morally wrong has been done, then someone did wrong.

( $H_3$ ) Neither Ann nor Ben did wrong.

MAC has no problem with rejecting  $H_3$ . The first acting agent did wrong – and if both acted simultaneously, both acted wrongly.

And with that, we have come to the end of a long journey, dug through the PYRAMID, and can now say: MOAC in the form of MAC has mastered the CHALLENGE, in its strongest form (cf. Figure 8.11).



**Figure 8.11:** The solution space of the CHALLENGE as the PYRAMID. We have reached the bottom, which is actually a climax.



# Chapter 9

## Summary and Future Work

Central to this undertaking was a comprehensive exploration and resolution of the CHALLENGE OF COLLECTIVE ACTION (concisely termed as the CHALLENGE) through the lens of MAXIMIZING OBJECTIVE ACT-CONSEQUENTIALISM (or, shorter, just MOAC). I claim that, against all odds, I actually succeeded. As the project progressed, its scope expanded beyond my original expectations, taking a more technical and formalized direction than initially anticipated. Reflecting on the overall endeavor, I think at least four philosophical insights are worth underscoring.

Firstly, the CHALLENGE is best understood as a mosaic of intricacies rather than a singular, overarching challenge. Its various manifestations stem from the apparently salient disparity between expected and actual assessments in specific collective contexts (aka

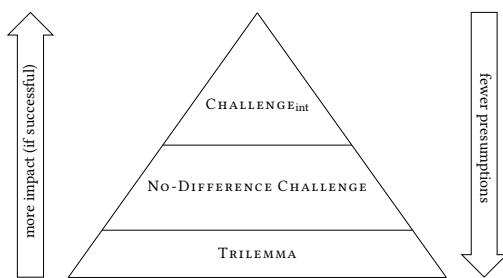


Figure 9.1: The three variants that make up the CHALLENGE as the PYRAMID as tackled in this thesis, in order of strength and number of preconditions they require. All layers have been solved.

TROUBLEMAKERS). Each variant has its own implications with different potential for devastation in the case of success, but they also come with quite different preconditions. ›The‹ strongest version of the CHALLENGE should thus be seen much more as a collection of mutually supportive challenges, where some backing others. My reconstruction of such a multi-layered version, the PYRAMID, consists of three such layers: The CHALLENGE as CHALLENGE<sub>int</sub>, the CHALLENGE as NO-DIFFERENCE CHALLENGE, and the CHALLENGE as TRILEMMA (cf. Figure 9.1).

Secondly, a more in-depth exploration reveals that CHALLENGE<sub>int</sub>, the most daunting yet assumption-laden variant of the CHALLENGE, in its most plausible interpretation, doesn't stand firm under scrutiny and must be dismissed as being invalid. This variant has garnered attention primarily because it spotlighted the implications of a reasoning trajectory that camp MOAC adopted so as not to fail collective scenarios straightforwardly. In reality, a careful investigation of MOAC indicates a discernible moral silence in a myriad of collective contexts, where MOCoR only allows to derive conditional assessments at best. This is the REAL CHALLENGE, the cause of all the collective struggle of MOAC theories. In an attempt to bridge these moral voids and to fill the deontic gaps badly, consequentialists expanded the evaluative contexts in collective decisions to include actions across all agents. Thus, refuting the gravest form of the CHALLENGE isn't a victory for consequentialism. It merely underscores that camp MOAC sidestepped a core issue and an even more fundamental problem, namely, the REAL CHALLENGE.

Fortunately, there was the third essential finding of my thesis, namely, that in this diagnosis, there also lies an opportunity for actual progress. It's indeed possible to precisely define proper consequences for each action by every in-

dividual agent within any collective decision situation. All that's required is the readiness to acknowledge decision situations as outcomes, mirroring the conventions of other decision-theoretic paradigms. This INTERMEDIATE OUTCOMES APPROACH (or simply the APPROACH) facilitates the breakdown of arbitrary collective decision situations into individual decision situations, ready to be crunched by MOCOR. By evaluating these newly discovered consequences, camp MOAC is essentially poised to bridge the deontic gaps in a better and less *ad hoc* way and thereby address not only the REAL CHALLENGE, but doing so in a way that does not reiterate the CHALLENGE.

Fourthly and finally, we refocused on the principle of PRINCIPLE OF MORAL HARMONY (or short PMH), which states that when multiple agents act in unison rightly, they are guaranteed to bring about the morally best outcome they collectively can bring about. Critical reflection revealed, however, that the prevalent formulations of PMH were overly stringent. Such definitions, in their commonly accepted breadth, directly logically and conceptually conflict with the foundational tenets of objective consequentialism. Recognizing this discordance, I introduced a more temperate alternative: the reasonable MORAL HARMONY (or simply, reasonable MH). With this new guideline, I navigated through various collective amendments that allowed camp MOAC to exploit the newly identified consequences. My journey culminated in adopting MIXED STRATEGIES, a theoretical amendment to the preceding MOAC framework. This finally brought us to MULTI-AGENT CONSEQUENTIALISM.

Throughout my research, I've introduced innovative tools and methodologies that further enrich the consequentialist realm. Notable among these are

the GENERALIZED EXTENSIVE FORM (GEF), which offers a comprehensive depiction of collective decision situations and the idea of theory-induced policies accompanied by their respective valuation.

## 9.1 Future Work?

While this thesis provides an in-depth exploration, it doesn't claim to be exhaustive. Various ideas, examples, and arguments were shelved, either due to deliberate limitations regarding the scope of this endeavor or due to their emergence late in the research process. What follows is an overview of potential avenues for future investigation, building on the foundation established in this work.

### 9.1.1 Implications for Subjective Consequentialism

There are, of course, different ideas about how objective and subjective variants of consequentialism relate to each other. As discussed in Section 2.3.1, especially in the setting of Principle 2.4, they can be viewed either as opposites or as complements. I tend to the second view and believe subjective consequentialist theories should set themselves the task of formulating decision procedures and possibly also criteria of rightness *>for real agents<* that should *approximate* the *>real<* objective criterion of rightness. Whether one shares this view or not, it exists in any case, and there is an implication for it from this work. Because it was spelled out here what is right according to MOAC (modulo axiological background theories) and how this is to be determined, the subjective theories understood as relaxations and approximation of that objective rightness predicate should be expected to be modified. It is necessary to work out these implications, and I have identified in this

idea the next normative ethical research project for me. This shall include a parameterization of subjective theories by different types of agents, including artificial ones. I think that this will pave the way to a machine ethics research program better rooted in philosophical theory.

### 9.1.2 Implications for The Actualism/Possibilism Debate

Recall the following part of A.N. Prior's quote from Section 6.5 (A. N. Prior and Raphael 1956, pp. 91-92):

Suppose that determinism is *not* true. Then there may indeed be a number of alternative actions which we could perform on a given occasion, but none of these actions can be said to have any »total consequences«, or to bring about a definite state of the world which is better than any other that might be brought about by other choices. For we may presume that other agents are free beside the one who is on the given occasion deciding what he ought to do, and the total future state of the world depends on how these others choose as well as on how the given person chooses ; and even if there were not other people to spoil one's calculations there would still be oneself, with one's own future choices, or some of them, undetermined like this present one (unless a man decides that it is too risky for him to have any further freewill, and on this very ground finds it to be his duty to do away with himself).

Prior undoubtedly has a point here when he highlights that one does not necessarily need other agents at all to raise some of the issues discussed in this book. In fact, the analogous challenge, in which the occurrence of certain consequences depends on what the same agent will do later and which is negotiated under the title »actualism/possibilism debate«, was already mentioned several times over the course of this project. The debate goes back at least to

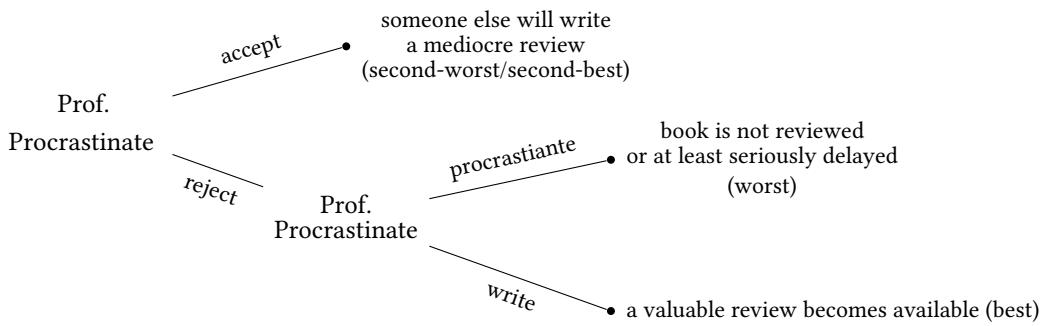
the late 1960s when Hector-Neri Castañeda Castaneda [1968] rightly observed that MOAC theories can often adequately assess sequences of actions according to their own criterion of rightness, but have problems inferring the rightness of individual actions.<sup>148</sup> Just as it was trivial for camp MOAC to name the best possible combinations of actions, it seemed difficult or even impossible to map this appropriately to the individual actions. This is, structurally speaking, the same issue, even though the agents (or, maybe more precisely, the temporal parts) involved are more intimately related. The *locus classicus* of modern times apparently remains an article by Frank Jackson and Robert Pargetter (cf. Jackson and Pargetter [1986]). Here is their infamous example (cf. *ibid.*, p. 235):

**Case 9.1 (PROFESSOR PROCRASTINATE)** *Professor Procrastinate receives an invitation to review a book. He is the best person to do the review, has the time, and so on. The best thing that can happen is that he says yes, and then writes the review when the book arrives. However, suppose it is further the case that were Procrastinate to say yes, he would not in fact get around to writing the review. Not because of incapacity or outside interference or anything like that, but because he would keep on putting the task off. (This has been known to happen.) Thus, although the best that can happen is for Procrastinate to say yes and then write, and he can do exactly this, what would in fact happen were he to say yes is that he would not write the review. Moreover, we may suppose, this latter is the worst that can happen. It would lead to the book not being reviewed at all, or at least to a review being seriously delayed.*

---

<sup>148</sup>I think that Chrisoula Andreou, in her take to the CHALLENGE (cf. Andreou [2014]), which according to private conversation, was never intended for consequentialists, is responding to a very similar observation.

Of course, we can represent this case as extensive form:



While PROFESSOR PROCRASTINATE is structurally similar to JOB MARKET and thus resembles a THRESHOLD CASES, there are also cases that seem to do without the temporal component and thus are more like COORDINATION CASES. Here is a case by Holly Goldman (Goldman 1978, p. 186, also cited by Jackson and Pargetter):

**Case 9.2 (Jones)** *Jones is driving through a tunnel behind a slow-moving truck. It is illegal to change lanes in the tunnel, and Jones's doing so would disrupt the traffic. Nevertheless, she is going to change lanes – perhaps she doesn't realize it is illegal, or perhaps she is simply in a hurry. If she changes lanes without accelerating, traffic will be disrupted more severely than if she accelerates. If she accelerates without changing lanes, her car will collide with the back of a truck.*

This decision situation can arguably be represented quite adequately in the following (in this case likely asymmetric) normal form:

JONES	accelerate	maintain speed
change lane	-	--
stay on lane	---	+

When Jones passes the truck, he should do it quickly. So the focus here is more on how something is done, and thus on the interplay of simultaneous actions – steering and accelerating – rather than on a sequence of actions (there is an obvious connection to what Douglas Portmore calls »Maximalism«, a position he connected to the PMH, cf. Portmore 2018).

It should be indisputable that there is a substantial structural similarity between those cases that dominate the actualism-possibilism debate and TROUBLEMAKERS. In the end, we seem to be dealing with almost the same challenge, only the agents involved are all one and the same agent. It is, therefore, reasonable to assume that if one commits oneself to a collective amendment, one also commits oneself to a position with respect to the actualism-possibilism debate for the sake of consistency. And indeed, when I set out to find comparable solutions, I found that Jackson himself proposed a solution quite similar to MIXED STRATEGIES and thus to MAC (cf. Jackson 2014).

And yet, I think there is still some work to be done here. After all, there is an interesting relationship, a kind of dependency between the agents in those cases that does not exist in TROUBLEMAKERS. Of course, this is not just *any* dependency but (diachronic) personal identity (or the closest thing to this relation). This may result in quite different possibilities and starting points. From investigating those possibilities, in interaction with the relaxation of the

independence condition crucial in the above analysis of collective decision situations, extensions of the approaches developed in this project could well arise. In this, an exciting project for future work can be discovered.

### 9.1.3 Generalization

I drove home many of my points against the backdrop of the simplest possible examples. I focused primarily on COORDINATION CASES, limiting myself as much as possible to cases with two agents with two cases each. I also excluded CUMULATIVE EFFECTS CASES (for what I think is a perfectly valid reason). Furthermore, I have not considered all possible collective amendments – and possibly not even all plausible ones. This project could be continued in terms of generalization along at least three dimensions.

*Explicit Generalization Through Inductive Reasoning* The cases I have discussed here can be regarded as *base cases* in the sense of formal, mathematical inductive reasoning. This is true for cases like Two FACTORIES, for cases like SEAMAN CLUMSY, and also for cases like JOB MARKET. But as the attentive reader will have noticed, my definitions are intentionally more general than they need to have been. For example, my definitions all allow that the consequences of individual actions in collective decision situations can themselves also be collective decision situations again. Although, of course, in the minimal cases, the collective decision situation reduced by action is an *individual* decision situation. (And thus are already ›termination states‹, representing final outcomes.)

In this respect, the *induction step* is already formally included, and to execute it is actually only drudgery (which, however, would surely again

cover several dozens, if not significantly more pages). I leave this task for future term papers and people who want to collect diligence points. Or for those who can discover errors when checking my lofty claims and use them for easy publications. (If, that is, anyone even takes note of this book.)

The thing becomes interesting, however, if one allows *infinities* (cf. Hedden [2020]). Both the case with infinitely many agents respectively actions – plausible if a moral universe could continue infinitely long, for instance – seems interesting to me, as well as the one with agents with infinitely many options. While for the first case, we should probably get quite far with discounting approaches (at least in cases of sequential action), the question is how to deal with other infinities in an orderly way. I would probably find a generalization in this direction more enjoyable than boring induction proofs.

*CUMULATIVE EFFECTS CASES* In section Section 4.2, I justified why I did not consider CUMULATIVE EFFECTS CASES in the context of this project. In essence, this is simply because there is sufficient doubt that these cases can be described coherently or consistently at all (cf. Kagan [2011]). Somewhat polemically simplified, how is the sum of many harms more harmful than ... the sum of harms?

I think the inhabitants of Camp MOAC may ask this question with raised eyebrows. At the same time, I think it can be made plausible, at least for some axiological theories (in the narrow and broader sense, cf. Section 2.3.3). Different arguments in this direction have different persuasiveness (cf. Nefsky [2011]; Spiekermann [2014]; Hedden [2020]). But even if, for example, we can provide a good argument for how the moral quality of outcomes can be unequal but on par, for example, in a hedonic, perceptual framework (cf. E. N. Dzhafarov

and D. D. Dzhafarov [2010a]; E. N. Dzhafarov and D. D. Dzhafarov [2010b]), consequentialists can probably respond to it by introducing renewed, matching aggregative amendments or by handling probabilistic, vague trigger cases by resort to probabilistic harms (cf. Shrader-Frechette [1987]).

I do not want to rule out addressing these questions in the future. I have some preliminary work on this in my ›Archive of the Unused Thousand Pages‹. But I am still waiting for the right moment. It will come at the latest when a suitable attack on MOAC is launched. Give me a proper attack, then I will deliver the counter. You don't have to shoot all your powder right away. For now, it is dry and well stored in the cellar of camp MOAC.

*More Sophisticated Amendments* Last but not least, one could also investigate further amendments. In my eyes, the most interesting ones would be those that try to include the fact that, often enough, not all agents act rightly. Maybe for an objective criterion of rightness, it is also perfectly okay to mark the reasonably best paths through our world full of imponderables. But perhaps, in the face of the newly accepted uncertainties stemming from METHODOLOGICAL INDETERMINISM, we should also make room for caution in the objective setting. Perhaps a *truly reasonable* formalization of PMH should accept more costs in the best case if, in turn, the more likely cases were better resolved. We would then abandon another idealizing assumption and, I guess, should again turn to decision theory to see what it offers us. I think it would be particularly promising to try out *regret-based* amendments (cf. Bell [1982]; Loomes and Sugden [1982]) and, more generally, to make the notion of *risk* fruitful for objective consequentialist theories. Further, I'd love to consider the details of an *ensemble solution* that allows one to choose between sev-

eral amendments (or decision rules) in dependence on certain features of the collective decision situation at hand (Chorus, Rose, and Hensher [2013]).

### 9.1.4 Formal Proofs

I claim that every major point I have made in this book is supported by argument. But, of course, not every argument is developed in detail. And every argument could be supported again by argument. Inevitably, this leads to, though probably not infinite, regress. This, of course, is not a weakness of *my* work but characteristic of (analytic) philosophical work.

But there could have been another way. This work has turned – twice – almost into an extension of various stit frameworks (Chellas [1992]; Horty and Belnap [1995]; Belnap, Perloff, and Xu [2001]; Horty [2001]). Then, in some semantic for multimodal logics, based on branching time models in the spirit of Prior (cf Arthur N. Prior [1955]; Arthur N. Prior [1967], I could have given formal proofs involving an act operator (or, to be true to tradition, a »seeing to it that« operator), a temporal operator, a modal operator, and a deontic operator. Perhaps I would have also saddled an epistemic operator on top, building on recent work by Horty and others (Horty [2019]; Ramírez Abarca and Broersen [2021]). In two years, I had written nearly 250 pages of this kind of stuff – and there was more than one fruitful and, I think, informative proof.

This material will certainly be published one day. The thing would become attractive if one could strengthen thereby, however, the points made here, based on natural language plus theorizing by those formal proofs. Because as long as one makes progress only within a formal system, one also proves only propositions within the system. For the matter to get philosophically really substantive, one would have to justify to what extent findings from the

formal system could be transferred to reality – and to what extent not.

Because I wanted to say something substantial about the world and not only (or primarily) about – or within – stit-semantics, I decided *not* to complete that project, returning to this earlier approach presented in this book. But I believe that based on first principles, on statements raised to the status of something like axioms for the description of relevant parts of reality, and if only against the backdrop of consequentialist theorizing, a sufficiently strong relation to a formal stit-systems could be established. I am sure that such a project is ultimately worth pursuing and definitely belongs in the corpus of future work.

## 9.2 How the Tables Have Turned

The CHALLENGE has so far been regarded as *particularly* delicate for act-consequentialism for two primary reasons: first, because the commonly diagnosed inability to assess actions that in combination lead to morally suboptimal consequences hits consequentialist basic intuitions; and second, to use Kagan’s phrase, because »consequentialism appears to fail even in its own favored terrain, where we are concerned with consequences and nothing but consequences.« I argue that my approach fundamentally underlines that the adequate processing of collective decision situations shall no longer seen as a weakness but, on the contrary, as a significant strength of consequentialism. Camp consequentialism now has a straightforward answer to the challenges of the complex interplay of many agents in mutual interdependence.

This matters also with respect to ›the battle of the families of moral theories‹. After all, camp deontology has struggles to deal with multi-agent scenar-

ios as well, especially with respect to overdetermination. As David Killoren and Bekka Williams put it (cf. Killoren and Bekka Williams [2013] p. 297):

Moreover, the *scope* of overdetermination problems seems to remain underappreciated. It is often assumed that overdetermination is mainly a problem for act-utilitarians (and other maximizing consequentialists), and that we are able to avoid such problems simply by shifting to a non-consequentialist view. That assumption is mistaken.

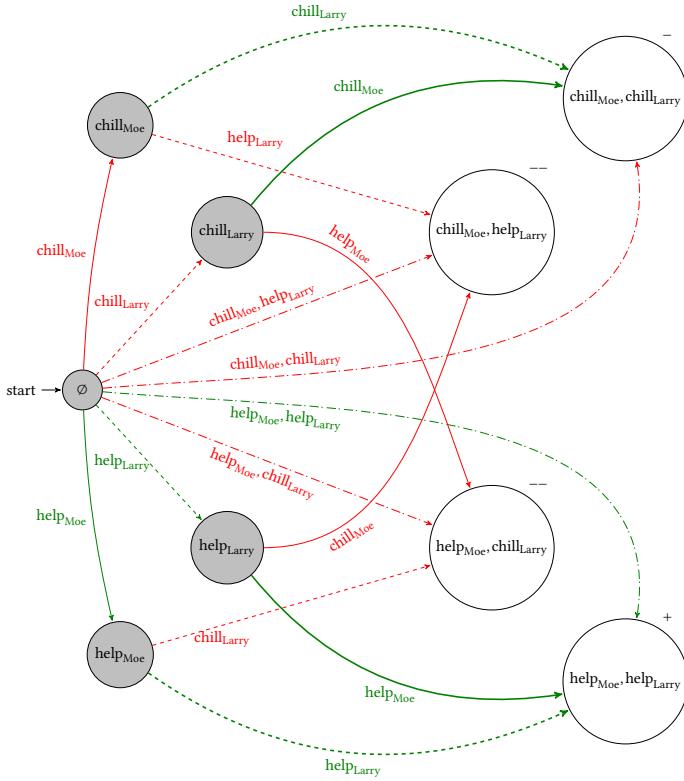
I couldn't put it better. Here is their example case meant to support their claim (cf. [ibid.](#), p. 297):

**Case 9.3 (STOOGES)** *Moe and Larry have promised to carry a piano upstairs by noon. This is a two-person job (neither stooge can carry the piano alone), and will require a half hour of time. Suppose that fulfilling this promise would result in greater utility than violating it [. . .] However, it is now 11:29 am, and both stooges are simply relaxing on the couch, each too lazy to put forth any effort.*

Here is, one last time, a normal form representing this case:

		Larry	
		chill	help
		Moe	
Moe	chill	promise goes unsatisfied, but both can chill (-)	promise goes unsatisfied, and Larry puts effort toward carrying the piano (--)
	help	promise goes unsatisfied, and Moe puts effort toward carrying the piano (--)	promise is fulfilled (+)

MAC can handle this case well. It is just another TROUBLEMAKER and we have derived what MIXED STRATEGIES plus MOCOR (EU) tells



**Figure 9.2:** The GEF of STOOGES with annotations according to MAC. We can observe how the combination of both, Larry and Moe, fulfilling their promise and helping move the piano is the right thing for them to do in the initial state.

Moe and Larry that it is right for them to hold their promise. Thus, act-consequentialism performs better (now) than Killoren and Williams suggest.

Even if we make this a *true* overdetermination problem, in which we assume that it is fixed that they will not show up to resolve their promise, the consequentialist has no problem resolving this situation appropriately. We just have to disambiguate what Killoren and Williams ask as to assume, namely: »They have decided to let noon pass without even touching the piano«). There are three cases to distinguish to consider: Either they colluded (in which case this collusion was wrong according to all we know, so there is an upstream wrong action); or one decided first not to go, in which case this first actor acted wrongly (for this we can have a look at the last GEF in

this book, see Figure 9.2); or both decide independently not to show up (in which case both acted wrongly, see also Figure 9.2).

I think Killoren and Williams diagnose the challenge correctly when they write (cf. Killoren and Bekka Williams 2013, p. 297):

[T]he breaking of the stooges' promise is ensured by the fact that Moe refuses to keep it, and is ensured by the fact that Larry refuses to keep it. Thus, neither stooge is individually able to bring it about that their promise is kept. [...]

Neither Moe nor Larry is able to keep their promise, given that both Moe and Larry are unwilling to do so. Thus, it is unclear whether we can justifiably say that either of them is (as an individual) morally obligated to do so.

But what is it that the non-consequentialists, especially the deontologists, are going to tell us? Killoren and Williams vote for an approach much like Jackson's (cf. Jackson 1987, see also Section 4.4.3), but without embracing Jackson's DIFFERENCE PRINCIPLE (i.e., Principle 3.1 from Section 3.3), allowing them to differ from Jackson's analysis with respect to the relation between the obligations of the individuals and their constructed (or inferred) group agent. However, we have seen such an approach, in general, means challenging METHODOLOGICAL INDIVIDUALISM, which is not a comfortable position to be in.<sup>149</sup>

I think, thus, the CHALLENGE remains unsolved for and unappreciated by camp deontology. Thus, the consequentialist has an advantage over the deontologist in the field of multi-agent scenarios – which, if we are honest, are ubiquitous. Ultimately, I posit that the most promising avenue for deon-

---

<sup>149</sup>Killoren and Williams are aware of this issue with their approach but are willing to bite the bullet (cf. Killoren and Bekka Williams 2013, p. 304): »But we are willing to grant that [Moe, Larry] fails to exhibit such behavior, given its rampant irrationality. [...] Despite all this, we want to call [Moe, Larry] a moral agent that has various obligations.«

tologists lies in the *consequentialization* of their theories (refer to Portmore 2009; Hurley 2020; Dreier 2011). Such a pivot equips them to leverage the APPROACH effectively. Once armed with these newly unearthed consequences, they may try to infer rights, obligations, and more as they see fit. That would be more honest, I claim provocatively, than to suggest group agents, where there are none, or to postulate *ad hoc* duties to cooperate, where cooperation was actually just excluded by assumption. You are always welcome here at camp Consequentialism! We are not afraid of more agents (not anymore!) – even if they are moral philosophers.



# Bibliography

- Andreou, Chrisoula (June 2014). »The Good, the Bad, and the Trivial.« In: *Philosophical Studies* 169.2, pp. 209–225 (cit. on pp. 4, 384).
- (Aug. 9, 2019). »Can Every Option Be Rationally Impermissible?« In: *Erkenntnis* (cit. on p. 179).
- Andrić, Vuko (2013). »Objective Consequentialism and the Licensing Dilemma.« In: *Philosophical Studies* 162.3, pp. 547–566 (cit. on p. 43).
- Bacharach, Michael (1999). »Interactive team reasoning: A contribution to the theory of co-operation.« In: *Research in economics* 53.2, pp. 117–147 (cit. on pp. 109, 259).
- (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press (cit. on p. 109).
- Baier, Kurt (1958). *The Moral Point of View a Rational Basis of Ethics*. Cornell University Press (cit. on pp. ii, iii, 115).
- Bell, David E. (1982). »Regret in Decision Making under Uncertainty.« In: *Operations Research* 30.5, pp. 961–981 (cit. on p. 389).
- Belnap, Nuel D., Michael Perloff, and Ming Xu (2001). *Facing the future: agents and choices in our indeterminist world*. Oxford University Press on Demand (cit. on pp. ix, 23, 390).

- Bentham, Jeremy (1780). *An Introduction to the Principles of Morals and Legislation*. Dover Publications (cit. on pp. [ii](#), [iii](#), [9](#), [115](#)).
- Berkeley, George (1712). »Passive Obedience.« In: pp. 427–469 (cit. on p. [114](#)).
- Blyth, Colin R. (1972). »On Simpson's Paradox and the Sure-Thing Principle.« In: *Journal of the American Statistical Association* 67.338, pp. 364–366 (cit. on p. [230](#)).
- Bolander, Thomas (2017). »Self-Reference.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University (cit. on p. [174](#)).
- Braham, Matthew and Martin van Hees (2011). »Responsibility voids.« In: *The Philosophical Quarterly* 61.242, pp. 6–15 (cit. on p. [127](#)).
- Brandt, Richard B. (1959). *Ethical Theory*. Englewood Cliffs, N.J.: Prentice-Hall (cit. on pp. [2](#), [76](#), [89](#)).
- Broad, C. D. (1916). »On the Function of False Hypotheses in Ethics.« In: *The International Journal of Ethics* 26.3, pp. 377–397 (cit. on pp. [75](#), [98](#)–[100](#), [102](#), [103](#)).
- Broome, John (2021). »How Much Harm Does Each of Us Do?« In: *Philosophy and Climate Change*. Ed. by Mark Budolfson, Tristram McPherson, and David Plunkett. Oxford University Press (cit. on p. [179](#)).
- Brown, Campbell (2011). »Consequentialize This.« In: *Ethics* 121.4, pp. 749–771 (cit. on pp. [128](#), [168](#)).
- Budolfson, Mark (2019). »The Inefficacy Objection to Consequentialism and the Problem with the Expected Consequences Response.« In: *Philosophical Studies* 176.7, pp. 1711–1724 (cit. on p. [4](#)).
- Budolfson, Mark, Tristram McPherson, and David Plunkett, eds. (2021). *Philosophy and Climate Change*. Oxford University Press (cit. on p. [vii](#)).

- Bykvist, Krister (2002). »Alternative Actions and the Spirit of Consequentialism.« In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 107.1, pp. 45–68 (cit. on p. 337).
- (2003). »Normative Supervenience and Consequentialism.« In: *Utilitas* 15.1, pp. 27–49 (cit. on pp. 29, 337).
- Castaneda, Hector-Neri (1968). »A Problem for Utilitarianism.« In: *Analysis* 28.4, pp. 141–142 (cit. on p. 384).
- Castañeda, Hector-Neri (1974). *The Structure of Morality*. Springfield, Ill., Thomas (cit. on pp. ii, iii, 115).
- Chellas, Brian F. (1992). »Time and Modality in the Logic of Agency.« In: *Studia Logica: An International Journal for Symbolic Logic* 51.3/4, pp. 485–517 (cit. on p. 390).
- Cholbi, Michael (2009). »The Murderer at the Door: What Kant Should Have Said.« In: *Philosophy and Phenomenological Research* 79.1, pp. 17–46 (cit. on p. 168).
- Chorus, Caspar G, John M Rose, and David A Hensher (2013). »Regret Minimization or Utility Maximization: It Depends on the Attribute.« In: *Environment and Planning B: Planning and Design* 40.1, pp. 154–169 (cit. on p. 390).
- Cohen, Daniel H. (1987). »The Problem of Counterpossibles.« In: *Notre Dame Journal of Formal Logic* 29.1, pp. 91–101 (cit. on p. 122).
- Dietz, Alexander (2020). »Are My Temporal Parts Agents?« In: *Philosophy and Phenomenological Research* 100.2, pp. 362–379 (cit. on p. 204).
- Dong, Xibin et al. (2020). »A Survey on Ensemble Learning.« In: *Frontiers of Computer Science* 14, pp. 241–258 (cit. on p. 369).

- Donne, John (1923). *Donne's Devotions*. Cambridge University Press (cit. on pp. [11], [290]).
- Dreier, James (1993). »Structures of Normative Theories.« In: *The Monist* 76.1, pp. 22–40 (cit. on p. [128]).
- (2011). »In Defense of Consequentializing.« In: *Oxford Studies in Normative Ethics, Volume 1*. Ed. by Mark Timmons. Oxford University Press (cit. on pp. [69], [128], [395]).
- Dzhafarov, Ehtibar N. and Damir D. Dzhafarov (2010a). »Sorites Without Vagueness I: Classificatory Sorites.« In: *Theoria* 76.1, pp. 4–24 (cit. on pp. [161], [388]).
- (2010b). »Sorites Without Vagueness II: Comparative Sorites.« In: *Theoria* 76.1, pp. 25–53 (cit. on pp. [90], [161], [389]).
- Estlund, David (Apr. 27, 2017). »Prime Justice.« In: *Political Utopias*. Ed. by Michael Weber and Kevin Vallier. Oxford University Press, pp. 35–56 (cit. on pp. [62], [63], [66], [68], [136], [137]).
- Fehige, Christoph (1995). »Das große Unglück der kleineren Zahl.« In: *Zum moralischen Denken*. Ed. by Christoph Fehige and Georg Meggle. Vol. 2. Suhrkamp. Chap. 8, pp. 139–175 (cit. on p. [50]).
- Feldman, Fred (1980). »The Principle of Moral Harmony.« In: *The Journal of Philosophy* 77.3, pp. 166–179 (cit. on pp. [ii], [iii], [4], [8], [9], [106], [114], [143], [179], [237], [249]).
- Ferreira, Jorge Viterbo and Franz Berto (2018). »The Problem of Counterpossibles.« In: (cit. on p. [122]).
- Gardiner, Stephen M (2011). *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford University Press (cit. on p. [vii]).

- Gibbard, Allan F. (1965). »Rule-Utilitarianism: Merely an Illusory Alternative?« In: *Australasian Journal of Philosophy* 43.2, pp. 211–220 (cit. on pp. 76, 109, 111).
- Glover, Jonathan and M. Scott-Taggart (1975). »“It Makes No Difference Whether or Not I Do It”.« In: *Aristotelian Society Supplementary Volume* 49.1, pp. 171–210 (cit. on pp. ii, iii, 4, 75, 76, 91, 157, 179, 186, 323).
- Gold, Natalie and Andrew M. Colman (2020). »Team Reasoning and the Rational Choice of Payoff-Dominant Outcomes in Games.« In: *Topoi* 39.2, pp. 305–316 (cit. on p. 109).
- Goldman, Holly S. (1978). »Doing the Best One Can.« In: *Values and Morals*. Ed. by Alvin Goldman and Jaegwon Kim. Reidel, pp. 185–214 (cit. on p. 385).
- Gori, Marco, Alessandro Betti, and Stefano Melacci (2023). *Machine Learning: A constraint-based approach*. Elsevier (cit. on p. 169).
- Gustafsson, Johan E. (2021). »Utilitarianism without Moral Aggregation.« In: *Canadian Journal of Philosophy* 51.4, pp. 256–269 (cit. on p. 47).
- Gustavsson, Kent (2021). »Charlie Dunbar Broad.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University (cit. on p. 106).
- Harrod, R. F. (1936). »Utilitarianism Revised.« In: *Mind* 45.178, pp. 137–156 (cit. on p. 76).
- Heath, Joseph (2020). »Methodological Individualism.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab, Stanford University (cit. on p. 32).
- Heathwood, Chris (2020). »An Opinionated Guide to "What Makes Someone's Life Go Best".« In: *Derek Parfit's Reasons and Persons: An Introduction and*

- Critical Inquiry*. Ed. by Andrea Sauchelli. Routledge, pp. 94–113 (cit. on p. 46).
- Hedden, Brian (2020). »Consequentialism and Collective Action.« In: *Ethics* 130.4, pp. 530–554 (cit. on pp. 4, 75, 120, 160, 161, 183, 388).
- Hooker, Brad (2023). »Rule Consequentialism.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University (cit. on p. 2).
- Horty, John F. (2001). *Agency and Deontic Logic*. Oxford University Press (cit. on pp. ix, x, 23, 29, 50, 237, 252, 289, 317, 390).
- (2019). »Epistemic Oughts in Stit Semantics.« In: *Ergo* 6.4, pp. 71–120 (cit. on pp. x, 390).
- Horty, John F. and Nuel D. Belnap (1995). »The Deliberative stit: A Study of Action, Omission, Ability, and Obligation.« In: *Journal of philosophical logic* 24.6, pp. 583–644 (cit. on pp. ix, 23, 390).
- Hurley, Paul (2020). »Consequentializing.« In: ed. by Douglas W. Portmore. Oxford University Press. Chap. 2, pp. 25–44 (cit. on p. 395).
- Jackson, Frank (1987). »Group Morality.« In: *Metaphysics and morality: essays in honour of J.J.C. Smart*. Ed. by J. J. C. Smart et al. Oxford, UK ; New York, NY, USA: B. Blackwell. Chap. 1, pp. 1–5 (cit. on pp. 4, 34, 66, 69, 80, 86, 101, 142, 178, 179, 186, 188–190, 394).
- (Apr. 1991). »Decision-Theoretic Consequentialism and the Nearest and Dearest Objection.« In: *Ethics* 101.3, pp. 461–482 (cit. on pp. 42, 48, 118, 286).
  - (1997). »Which Effects.« In: *Reading Parfit*. Ed. by J. Dancy. Blackwell, pp. 42–53 (cit. on pp. 80, 86).

- (2014). »Procrastinate Revisited.« In: *Pacific Philosophical Quarterly* 95.4, pp. 634–647 (cit. on pp. 204, 289, 386).
- Jackson, Frank and Robert Pargetter (Apr. 1986). »Oughts, Options, and Actuality.« In: *The Philosophical Review* 95.2, p. 233 (cit. on pp. 204, 384).
- Jeffrey, Richard (1982). »The Sure Thing Principle.« In: *PSA: Proceedings of the biennial meeting of the philosophy of science association*. Vol. 1982. 2. Philosophy of Science Association, pp. 719–730 (cit. on p. 230).
- Kagan, Shelly (2011). »Do I Make a Difference?: Do I Make a Difference?« In: *Philosophy & Public Affairs* 39.2, pp. 105–141 (cit. on pp. ii, iii, 4, 13, 74, 80, 83, 90, 92, 94, 120, 143, 159, 178–182, 288, 388).
- Kavka, Gregory S (1983). »The Toxin Puzzle.« In: *Analysis* 43.1, pp. 33–36 (cit. on p. 40).
- Kemeny, J.G. and J.L. Snell (1960). *Finite Markov Chains*. Finite Markov Chains. Van Nostrand (cit. on p. 293).
- Killoren, David and Bekka Williams (Apr. 2013). »Group Agency and Overdetermination.« In: *Ethical Theory and Moral Practice* 16.2, pp. 295–307 (cit. on pp. 4, 392, 394).
- Lawford-Smith, Holly and William Tuckwell (2020). »Act Consequentialism and the No-Difference Challenge.« In: ed. by Douglas W. Portmore. Oxford University Press. Chap. 33, pp. 634–654 (cit. on pp. 72, 75).
- Lewis, C. I. (1932). »Alternative Systems of Logic.« In: *The Monist* 42.4, pp. 481–507 (cit. on p. 132).
- Lewis, David (1973). *Counterfactuals*. Blackwell (cit. on p. 122).
- List, Christian and Philip Pettit (2011). *Group Agency: The Possibility, Design, Status of Corporate Agents*. Oxford: Oxford Univ. Press. 238 pp. (cit. on pp. 21, 34).

- Loomes, Graham and Robert Sugden (1982). »Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty.« In: *The Economic Journal* 92.368, pp. 805–824 (cit. on p. [389]).
- MacFarlane, John (2023). »Belief: What is it Good for?« In: *Erkenntnis* (cit. on p. [43]).
- Mackie, John Leslie (1977). *Ethics: Inventing Right and Wrong*. Penguin Books (cit. on pp. [ii], [iii], [115]).
- McCall, Storrs (1973). In: *Synthese* 26.1, pp. 165–171 (cit. on p. [132]).
- McNamara, Paul and Frederik Van De Putte (2022). »Deontic Logic.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University (cit. on p. [30]).
- Moore, George Edward (1903). *Principia Ethica*. Ed. by Thomas Baldwin. Mineola, N.Y.: Dover Publications (cit. on p. [252]).
- Nash, John (1950). »The Bargaining Problem.« In: *Econometrica* 18.2, pp. 155–162 (cit. on p. [322]).
- (1951). »Non-Cooperative Games.« In: *Annals of Mathematics* 54.2, pp. 286–295 (cit. on pp. [141], [322]).
- Nefsky, Julia (2011). »Consequentialism and the Problem of Collective Harm: A Reply to Kagan.« In: *Philosophy and Public Affairs* 39.4, pp. 364–395 (cit. on pp. [4], [90], [160], [388]).
- Notz, Dirk and Julienne Stroeve (2016). »Observed Arctic Sea-ice Loss Directly Follows Anthropogenic CO<sub>2</sub> Emission.« In: *Science* 354.6313, pp. 747–750 (cit. on p. [157]).
- Ord, Toby (2005). »Consequentialism and Decision Procedures.« PhD thesis. University of Oxford (cit. on p. [43]).

- Parfit, Derek (1984). *Reasons and Persons*. Oxford University Press (cit. on pp. ii, iii, 4, 46, 80, 90, 94, 109, 142, 158, 179, 186, 225).
- (1988). »What We Together Do.« In: pp. 1–33 (cit. on pp. 4, 43, 91, 92, 118, 142, 179, 186, 259).
- Pearl, Judea et al. (2000). »Models, Reasoning and Inference.« In: *Cambridge, UK: Cambridge University Press* 19 (cit. on p. 230).
- Pearl, Judea (2016). »The Sure-Thing Principle.« In: *Journal of Causal Inference* 4.1, pp. 81–86 (cit. on p. 230).
- Peirce, Charles Sanders (1931). *Collected Papers*. Cambridge, MA, USA: Harvard University Press (cit. on p. 151).
- Petersson, Björn (2017). »Team Reasoning and Collective Intentionality.« In: *Review of Philosophy and Psychology* 8.2, pp. 199–218 (cit. on p. 109).
- Pettit, Philip and David Schweikard (Mar. 2006). »Joint Actions and Group Agents.« In: *Philosophy of the Social Sciences* 36.1, pp. 18–39 (cit. on p. 34).
- Pinkert, Felix (2015). »What If I Cannot Make a Difference (and Know It).« In: *Ethics* 125.4, pp. 971–998 (cit. on pp. ii, iii, 4, 8, 24, 63, 87, 109, 116, 117, 125, 138, 178, 179, 183, 185, 267, 292).
- Portmore, Douglas W. (2007). »Consequentializing Moral Theories.« In: *Pacific Philosophical Quarterly* 88.1, pp. 39–73 (cit. on p. 69).
- (2009). »Consequentializing.« In: *Philosophy Compass* 4.2, pp. 329–347 (cit. on pp. 69, 128, 395).
- (2018). »Maximalism and Moral Harmony.« In: *Philosophy and Phenomenological Research* 96.2, pp. 318–341 (cit. on pp. 4, 8, 9, 108, 109, 115, 121, 143, 179, 249, 386).
- ed. (2020). *The Oxford Handbook of Consequentialism*. Oxford University Press.

- Prior, A. N. and D. D. Raphael (1956). »The Consequences of Actions.« In: *Aristotelian Society Supplementary Volume* 30.1, pp. 91–119 (cit. on pp. 252, 289, 313, 383).
- Prior, Arthur N. (1955). *Time and Modality*. Greenwood Press (cit. on pp. x, 23, 390).
- (1967). »Past, present and future.« In: *Revue Philosophique de la France Et de l'Étranger* 157 (cit. on pp. x, 23, 390).
- Quinn, Warren (1993). »The Puzzle of the Self-Torturer.« In: *Morality and Action* (cit. on p. 73).
- Railton, Peter (1984). »Alienation, Consequentialism, and the Demands of Morality.« In: *Philosophy & Public Affairs* 13.2, pp. 134–171 (cit. on pp. 40, 119).
- Ramírez Abarca, Aldo Iván and Jan Broersen (2021). »Stit Semantics for Epistemic Notions Based on Information Disclosure in Interactive Settings.« In: *Journal of Logical and Algebraic Methods in Programming* 123, p. 100708 (cit. on p. 390).
- Rawls, John (1971). *A Theory of Justice*. Eleventh printing, 1981. Cambridge, Mass: The Belknap press of Harvard University Press (cit. on p. 319).
- Regan, Donald H. (1980). *Utilitarianism and Co-Operation*. Oxford University Press (cit. on pp. ii, iii, 4, 8, 76, 106, 108, 109, 111, 113, 118, 123–125, 129, 131, 132, 179, 186, 271–273, 275).
- Savage, Leonard J (1954). »The Foundations of Statistics.« In: (cit. on p. 230).
- Schelling, T.C. (1980). *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press (cit. on p. 109).
- Schnieder, Benjamin (2011). »A Logic for 'Because'.« In: *Review of Symbolic Logic* 4.3, pp. 445–465 (cit. on p. 191).

- Schroeder, Mark (2021). »Value Theory.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University (cit. on p. 46).
- Schumpeter, J.A. (1908). *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Duncker & Humblot (cit. on p. 32).
- Shrader-Frechette, Kristin (Oct. 1987). »Parfit and Mistakes in Moral Mathematics.« In: *Ethics* 98.1, pp. 50–60 (cit. on pp. 179, 389).
- Simpson, E. H. (1951). »The Interpretation of Interaction in Contingency Tables.« In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.2, pp. 238–241 (cit. on p. 230).
- Singer, Peter (1980). »Utilitarianism and Vegetarianism.« In: *Philosophy & Public Affairs* 9.4, pp. 325–337 (cit. on p. 120).
- Sinnott-Armstrong, Walter (2005). »It's Not My Fault: Global Warming and Individual Moral Obligations.« In: *Perspectives on Climate Change*. Ed. by Walter Sinnott-Armstrong and Richard Howarth. Elsevier, pp. 221–253 (cit. on p. 4).
- (2022). »Consequentialism.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University (cit. on pp. 40, 289).
- Sinnott-Armstrong, Walter and Richard B. Howarth, eds. (2005). *Perspectives on Climate Change: Science, Economics, Politics, Ethics*. 1. ed. Advances in the economics of environmental resources 5. OCLC: 254784373. Amsterdam: Elsevier JAI. 307 pp. (cit. on pp. viii, 5).
- Slote, Michael (1984). »Morality and Self-Other Asymmetry.« In: *The Journal of Philosophy* 81.4, pp. 179–192 (cit. on p. 73).

- Slote, Michael and Philip Pettit (1984). »Satisficing Consequentialism.« In: *Aristotelian Society Supplementary Volume* 58.1, pp. 139–176 (cit. on p. 2).
- Smart, J. J. C. (1973). »An Outline of a System of Utilitarian Ethics.« In: Cambridge [Eng.]: University Press (cit. on pp. 77, 179, 323, 324).
- Smart, J. J. C. and Bernard Williams (1973). *Utilitarianism: For and Against*. Cambridge [Eng.]: University Press. 155 pp. (cit. on p. 75).
- Spiekermann, Kai (Sept. 2014). »Causing Harm with Others: Small Impacts and Imperceptible Effects.« In: *Midwest Studies In Philosophy* 38.1, pp. 75–90 (cit. on pp. 86, 160, 161, 388).
- Sverdlik, Steven (2011). *Motive and Rightness*. Oxford, GB: Oxford University Press UK (cit. on p. 2).
- Textor, Mark (2021). »States of Affairs.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cit. on p. 23).
- Timmerman, Travis and Yishai Cohen (2020). »Actualism and Possibilism in Ethics.« In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University (cit. on p. 204).
- Timmons, Mark (2001). *Moral Theory: An Introduction*. Lanham, Md.: Rowman & Littlefield Publishers (cit. on pp. 32, 44, 164, 167).
- Toulmin, Stephen E. (1953). *An examination of the place of reason in ethics*. Cambridge at the University Press (cit. on pp. 115, 249).
- von Neumann, John and Oskar Morgenstern (1947). *Theory of Games and Economic Behavior*. Second. New Jersey: Princeton University Press (cit. on p. 48).

- Weber, M., G. Roth, and C. Wittich (1978). *Economy and Society: An Outline of Interpretive Sociology*. Economy and Society: An Outline of Interpretative Sociology. University of California Press (cit. on p. [32]).
- Williamson, Timothy (Oct. 2002). *Knowledge and its Limits*. Oxford University Press (cit. on p. [43]).
- Woodard, C. (Apr. 1, 2009). »What's Wrong with Possibilism.« In: *Analysis* 69.2, pp. 219–226 (cit. on p. [204]).
- Woodard, Christopher (2003). »Group-Based Reasons for Action.« In: *Ethical Theory and Moral Practice* 6.2, pp. 215–229 (cit. on p. [63]).
- (Sept. 2019). *Taking Utilitarianism Seriously*. Oxford University Press (cit. on p. [9]).
- Wroński, Leszek (2020). »Objective Consequentialism and the Plurality of Chances.« In: *Synthese* 198.12, pp. 12089–12105 (cit. on p. [289]).
- Zamir, T. (2001). »One Consequence of Consequentialism: Morality and Overdetermination.« In: *Erkenntnis* 55.2, pp. 155–168 (cit. on pp. [80], [86]).
- Zimmerman, Michael J. (1996). *The Concept of Moral Obligation*. Cambridge University Press (cit. on pp. [ii], [iii], [4], [68], [88], [115], [140], [142], [179]).
- (2014). *Ignorance and Moral Obligation*. Oxford University Press (cit. on p. [41]).

**Used AI Tools and the Extent of Their Use:** Three AI-supported applications were used in the creation of this dissertation: DeepL for translation suggestions, ChatGPT (especially the version based on GPT 4) for reformulation suggestions and the Grammarly service as a plugin for Overleaf in Safari to avoid the worst grammatical errors. Importantly, ChatGPT was only used to formulate my own thoughts, which I submitted via prompt. All text generated in this way was checked by me, incorporated into the text independently, and adapted to the specific context-dependent requirements. No text blocks were adopted in their entirety.



# What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research <sup>☆,☆☆</sup>



Markus Langer <sup>a,\*<sup>1</sup></sup>, Daniel Oster <sup>b,1</sup>, Timo Speith <sup>b,c,\*<sup>1</sup></sup>, Holger Hermanns <sup>c,d</sup>, Lena Kästner <sup>b</sup>, Eva Schmidt <sup>e</sup>, Andreas Sesan <sup>f</sup>, Kevin Baum <sup>b,c</sup>

<sup>a</sup> Department of Psychology, Saarland University, Saarbrücken, Germany

<sup>b</sup> Institute of Philosophy, Saarland University, Saarbrücken, Germany

<sup>c</sup> Department of Computer Science, Saarland University, Saarbrücken, Germany

<sup>d</sup> Institute of Intelligent Software, Guangzhou, China

<sup>e</sup> Institute of Philosophy and Political Sciences, Technical University Dortmund, Germany

<sup>f</sup> Institute of Legal Informatics, Saarland University, Saarbrücken, Germany

## ARTICLE INFO

### Article history:

Received 1 May 2020

Received in revised form 8 February 2021

Accepted 9 February 2021

Available online 15 February 2021

### Keywords:

Explainable Artificial Intelligence

Explainability

Interpretability

Explanations

Understanding

Interdisciplinary Research

Human-Computer Interaction

## ABSTRACT

Previous research in Explainable Artificial Intelligence (XAI) suggests that a main aim of explainability approaches is to satisfy specific interests, goals, expectations, needs, and demands regarding artificial systems (we call these “*stakeholders’ desiderata*”) in a variety of contexts. However, the literature on XAI is vast, spreads out across multiple largely disconnected disciplines, and it often remains unclear *how* explainability approaches are supposed to achieve the goal of satisfying stakeholders’ desiderata. This paper discusses the main classes of stakeholders calling for explainability of artificial systems and reviews their desiderata. We provide a model that explicitly spells out the main concepts and relations necessary to consider and investigate when evaluating, adjusting, choosing, and developing explainability approaches that aim to satisfy stakeholders’ desiderata. This model can serve researchers from the variety of different disciplines involved in XAI as a common ground. It emphasizes where there is interdisciplinary potential in the evaluation and the development of explainability approaches.

© 2021 Elsevier B.V. All rights reserved.

## 1. Introduction

### 1.1. Background, motivation, and related work

Explainable Artificial Intelligence (XAI) is – once again [1–4] – a burgeoning multidisciplinary area of research. In general, XAI can be perceived as the topic or research field concerned with developing approaches to explain and make artificial systems understandable to human stakeholders [5–7].

<sup>☆</sup> Work on this paper was funded by the Volkswagen Foundation grants AZ 95143, AZ 98509, AZ 98510, AZ 98511, AZ 98512, AZ 98513, and AZ 98514 “Explainable Intelligent Systems” (EIS), by the DFG grant 389792660 as part of TRR 248, and by the ERC Advanced Grant 695614 (POWVER). The authors thank the anonymous reviewers for feedback on initial drafts of this paper.

<sup>☆☆</sup> This paper is part of the Special Issue on Explainable AI.

\* Corresponding authors.

E-mail address: [timo.speith@uni-saarland.de](mailto:timo.speith@uni-saarland.de) (T. Speith).

<sup>1</sup> Markus Langer, Daniel Oster, and Timo Speith have contributed equally to this article and share the first authorship.

This puts several central aspects into the focus of XAI research. First, artificial systems are the primary objects of investigation. Such systems can range from systems following a predefined set of rules, to expert and knowledge-based systems, to systems relying on machine learning. Insights from XAI research become important when these systems are too complex to allow for human oversight or are inherently opaque, which precludes human insight [8]. Second, this view on XAI emphasizes the importance of approaches that enable or provide insights into artificial systems, their functioning, and their outputs. These approaches (we call them '*explainability approaches*') encompass methods, procedures, and strategies to provide explanatory information helping us to better understand artificial systems. Third, there is a decisive need for XAI because there are human stakeholders (e.g., users, developers, regulators)<sup>2</sup> whose interests, goals, expectations, needs, and demands regarding artificial systems (e.g., to have fair or trustworthy systems [9,10]) call for greater understandability of artificial systems. We call such conglomerations of stakeholders' interests, goals, expectations, needs, and demands regarding artificial systems '*stakeholders' desiderata*'.

A large part of previous XAI research was mainly concerned with developing new explainability approaches without evaluating whether these methods are useful to satisfy stakeholders' desiderata (except maybe the desiderata of developers) [9,11–13]. In fact, only a minority of papers concerned with explainability approaches also evaluated the proposed methods [12,13]. In contrast, nowadays an increasing number of researchers strongly suggest putting human stakeholders in the center of attention when evaluating and developing explainability approaches. For instance, researchers have proposed to comprehensively examine the perspectives of all stakeholders involved in the discussions around XAI (e.g., [14–16]) or they have introduced evaluation methods and metrics to systematically and empirically investigate the effects of explainability approaches on human stakeholders and their desiderata (e.g., [9,17]).

This paper reinforces and extends the focus on human stakeholders as well as on the development and evaluation of explainability approaches, and provides three main contributions. First, we propose that when evaluating, adjusting, choosing, and developing explainability approaches, research needs to pay more attention to stakeholders' specific desiderata in given contexts. This is crucial as the success of explainability approaches depends on how well they satisfy these desiderata. Current measures and metrics focus on how well explainability approaches calibrate trust or how much they increase human-machine performance (see, e.g., [17]). However, these are just two of many desiderata driving XAI research and although there is research that investigates which classes of stakeholders hold essential desiderata for XAI (e.g., [14,15,18,19]), there is a lack of research identifying, defining, and empirically investigating these desiderata, let alone research that links them to explainability approaches suitable for their satisfaction. Our paper identifies desiderata of different classes of stakeholders and calls for systematic empirical research investigating how explainability approaches, through the facilitation of understanding, lead to the satisfaction of these desiderata.

Second, we emphasize the central role of understanding as a path through which explainability approaches satisfy stakeholders' desiderata. Although research has highlighted the importance of human understanding to XAI (e.g., [17,20]), understanding sometimes seems to be considered as just one of many important outcomes of explainability approaches [21]. We claim that increasing human understanding is not just one of many important effects of explainability approaches, but crucial for the satisfaction of desiderata in general. For this reason, we introduce a model that emphasizes the critical importance of human understanding as a mediator between explainability approaches and the satisfaction of desiderata (for a related model focused on user performance, see Hoffman et al. [17]).

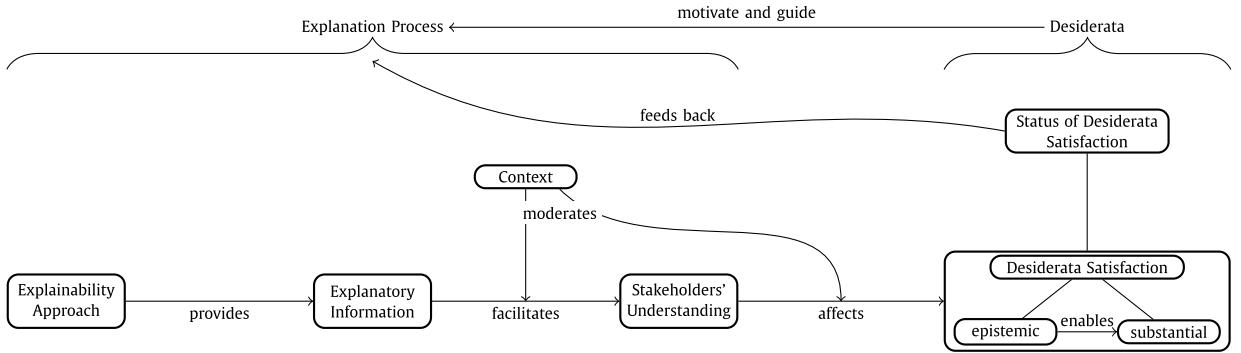
Third, we propose that our model can be used to guide evaluating, adjusting, choosing, and developing explainability approaches. In particular, our model highlights the main concepts and their relations of how explainability approaches are supposed to lead to the satisfaction of stakeholders' desiderata. Clearly defining, analyzing, and capturing these concepts, as well as clarifying their relations are central for the systematic evaluation of the success of explainability approaches, as this helps to identify potential reasons for why an explainability approach did not satisfy given desiderata. Similarly, considering these concepts and their relations is crucial for the choice between different explainability approaches or for the successful development of such approaches, because this supports the derivation of requirements for an explainability approach that has the potential to satisfy stakeholders' desiderata. Furthermore, our model is useful to detect where input from disciplines outside of computer science (e.g., psychology, philosophy, law, sociology; [22]) is crucial when evaluating or developing explainability approaches. Thus, our model serves to identify interdisciplinary potential and is aimed to establish a common ground for different disciplines involved in XAI. Overall, the current paper is intended for an interdisciplinary readership interested in XAI.

## 1.2. A conceptual model of the relation between explainability approaches and stakeholders' desiderata

For the purposes of this paper, we introduce a conceptual model (see Fig. 1) that organizes and makes explicit the central concepts of how explainability approaches relate to the satisfaction of stakeholders' desiderata, as well as the relations between these concepts. The main concepts in this model are: '*explainability approach*', '*explanatory information*', '*stakeholders' understanding*', '*desiderata satisfaction*', and '*(given) context*'.

The overall idea of our model is that the success of an explainability approach depends on the satisfaction of stakeholders' desiderata (consisting of the *substantial* and the *epistemic* facet of desiderata satisfaction, see Section 3.1). Desiderata

<sup>2</sup> Generally, we speak of single stakeholders here. Since we cannot consider each stakeholder individually, we treat them as representative members of specified stakeholder classes.



**Fig. 1.** Our proposed model of how explainability approaches relate to the satisfaction of stakeholders' desiderata.

satisfaction, thus, motivates an explanation process including explainability approaches, explanatory information, and stakeholders' understanding. Specifically, in the explanation process we assume that explainability approaches provide explanatory information to human stakeholders. Human stakeholders engage with the information to facilitate their understanding of an artificial system, its functioning, and outputs. As a consequence, the adjusted understanding of the stakeholders affects the extent to which their desiderata are satisfied. The context in which the human stakeholder and the artificial system operate and interact affects the relations between the other concepts (i.e., influences the relation between explanatory information and stakeholder understanding as well as the relation between understanding and desiderata satisfaction). Identifying, defining, as well as capturing and empirically examining the concepts and their relations should guide evaluating, adjusting, choosing, and developing of explainability approaches that aim to satisfy stakeholders' desiderata.

With a focus on stakeholders' desiderata, the following sections will elaborate on the model's concepts and their relations in more detail, as well as explicate shortcomings of the current view on these concepts and their relations. This paper is structured as follows. We will start on the right side of Fig. 1 and will continue to work backwards from stakeholders' desiderata. In Section 2, we will describe different classes of stakeholders and provide examples of their pertinent desiderata. We will elaborate on the central role of understanding for desiderata satisfaction in Section 3. Evoking understanding, in turn, requires explanatory information, as we will illuminate in Section 4. Section 5 will shed light on the connection between explainability approaches and explanatory information. Throughout these sections, we will point towards interdisciplinary potential that becomes apparent with the transition from one concept to the next in our model. In Section 6, we will, then, exemplify how our model can be used to evaluate, adjust, choose, and develop explainability approaches.

## 2. Stakeholders' desiderata

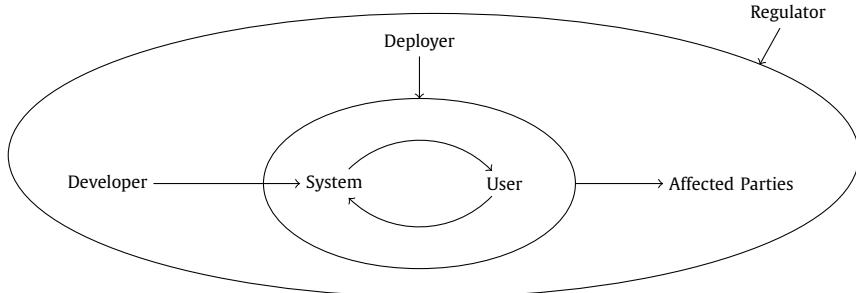
Stakeholders' desiderata are one, if not *the*, reason for the rising popularity of XAI (see also [14,19]). Since stakeholders in combination with their concrete desiderata motivate, guide, and affect the explanation process depicted in Fig. 1, we propose that identifying and clarifying desiderata of the various classes of stakeholders related to artificial systems is a crucial first step when evaluating, adjusting, choosing, and developing explainability approaches for an artificial system in a given context.

### 2.1. Stakeholder classes

The need for explainability starts with the increasing societal impact of artificial systems and the fact that many such systems still have to be operated by humans or affect human lives. This indicates that there are various groups of people with different interests in the explainability of artificial systems: people operate systems, try to improve them, are affected by decisions based on their output, deploy systems for everyday tasks, and set the regulatory frame for their use. These people are commonly called *stakeholders*.<sup>3</sup>

Previous research has discussed varying classes of stakeholders in the context of XAI. For instance, Preece et al. [18] distinguish between four main classes of stakeholders: *developers*, *theorists*, *ethicists*, and *users*. Arrieta et al. [14] categorize the main classes of stakeholders into domain experts/users, data scientists/developers/product owners, users affected by model decisions, managers/executive board members, and regulatory entities (see also [15,19,24]). We follow Arrieta et al. and distinguish five classes of stakeholders: *users*, (*system*) *developers*, *affected parties*, *deployers*, and *regulators* (see Fig. 2).

<sup>3</sup> According to the Merriam-Webster dictionary, a stakeholder is, among others, someone 'who is involved in or affected by a course of action' [23]. For this reason, we use this as a general term, but refer to specific stakeholder classes where appropriate.



**Fig. 2.** The different classes of stakeholders associated with artificial systems and their relations.

Clearly, one person can be a member of several stakeholder classes. A user, for instance, can be affected by the outputs of the system she operates. Additionally, these are just prototypical classes of stakeholders and more fine-grained distinctions into sub-classes of stakeholders are possible [15]. For example, there is not one prototypical developer, but developers differ in their expertise and in other factors (e.g., personality). A novice developer may have different desiderata than an expert. In a similar way, a lay user's desiderata might differ from those of an expert user (more on this in Section 4.2). Moreover, we want to emphasize that this list of stakeholders is not necessarily exhaustive because our distinction is based on previous research that mainly comes from a computer scientific background. Thus, it might neglect other classes of stakeholders.

## 2.2. Exemplary stakeholders' desiderata

The desiderata arising from the five classes of stakeholders are diverse. Based on a term search, we identified more than 100 peer-reviewed journal and conference publications that postulate XAI as being indispensable when it comes to satisfying the different desiderata (see Table 1).

Table 1 presents exemplary desiderata that we have extracted from this literature review. Each row contains a desideratum, partnered with sources that claim, propose, or show that XAI-related research (e.g., on explainability approaches) and its findings and outputs can contribute to the satisfaction of this desideratum. Furthermore, this table presents stakeholder classes that may be most prone to have one or more of these desiderata. Whenever we could not extract stakeholder classes from the respective papers, we did our own (mostly common-sense) mappings. Notably, most of the sources we present in this table only claim that XAI-related research can contribute to satisfy the respective desiderata with only a subset of these papers (for instance, [28,31,45]) providing empirical evidence for their claims (e.g., regarding the mapping of desiderata to stakeholder classes or regarding the relation of explanatory information and desiderata). In what follows, we present two important exemplary desiderata for each class of stakeholders.

**Users.** Most papers concerning stakeholders in XAI have this class of stakeholders in common (see, e.g., [14,18,19]). Among others, users take into account recommendations of artificial systems to make decisions [24]. Some prototypical members of this stakeholder class are medical doctors, loan officers, judges, or hiring managers. Usually, users are no experts regarding the technical details and the functioning of the systems they use. However, they can work more effectively if they form adequate expectations concerning the systems' functioning. In case they cannot do so, and in cases where their expectations are violated, they need information that goes beyond the knowledge of purely operating the system. This motivates at least the following two central desiderata of users: *usability* [22,36] and *trust* [29,96,115].

In many cases, a system is more usable if it offers meaningful information alongside its outputs. This information can help users to adequately link their knowledge and assessment of a given situation to the information used by a system, can help them to make decisions more quickly, or to increase decision quality [116]. All of this can contribute to the usefulness (another important desideratum of users) of a system and is important in high-stakes scenarios where a user decides on the basis of a system's recommendations.

This is closely linked to the desideratum of adequately calibrating trust in systems. Both undertrust and overtrust can negatively affect the appropriate use of systems [117]. In the case of undertrust, users may constantly try to supervise a system's behavior or even attempt to intervene in a system's processes, thereby undermining the effectiveness of the human-system interaction [118]. In the case of overtrust, people may use a system without questioning its behavior [118–120]. This can again decrease the effectiveness of the human-system interaction, as humans rely on the system's outputs even in situations where they should challenge them [117,121]. Explainability approaches have the potential to provide means to let users adequately calibrate their trust in artificial systems [17].

**Developers.** Individuals who design, program, and build artificial systems are the developers. Naturally, they count as a class of stakeholders, as without them the systems would not exist in the first place. Generally, developers have a high expertise concerning the systems and an interest in improving them.

An especially important desideratum of developers is *verification*, that is, to check whether a system works as intended [7,122–124]. There are many ML-based classifiers that consider, for instance, irrelevant inputs as relevant (see, e.g., [125,126]). Increasing insights into the system's decision-making processes by using certain explainability approaches can help

**Table 1**

An exemplary list of desiderata, stakeholders holding these desiderata, and sources that claim, propose, or show that XAI-related research (e.g., on explainability approaches) and its findings and outputs can contribute to the satisfaction of these desiderata.

Desideratum	Tentative description	Stakeholder	Without empirical (emp.) investigation	Emp. evidence	Emp. mixed	Emp. inconclusive
Acceptance	Improve acceptance of systems	Deployer, Regulator	[5,13,15,25–48]	[49]	[50]	
Accountability	Provide appropriate means to determine who is accountable	Regulator	[7,15,16,20,22,29,51–67]	[43,68]		
Accuracy	Assess and increase a system's predictive accuracy	Developer	[5,25,33,38,46,51,55,69–73]			
Autonomy	Enable humans to retain their autonomy when interacting with a system	User	[15,44,53,56,57,60,74–76]			
Confidence	Make humans confident when using a system	User	[5,14–16,28,32,34–36,38,39,48,49,57,58,69,73,74,77–81]			[82]
Controllability	Retain (complete) human control concerning a system	User	[6,12,13,22,44,49,56,60,74]	[39]	[43]	
Debugability	Identify and fix errors and bugs	Developer	[5,7,12,13,16,25,30,33,44,47,48,51,56,58,61–64,67,74,83–88]	[82,89,90]		
Education	Learn how to use a system and system's peculiarities	User	[6,13,24,31,33,36,39,49,74,80,88,91,92]			
Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	Developer, User	[13,26,34,36,39,42,48,49,65,67,74,77,78,93–95]	[28,31,46]		
Efficiency	Assess and increase a system's efficiency; work efficiently with a system	Developer, User	[13,28,39,55,74,78,93–95]	[31]		
Fairness	Assess and increase a system's (actual) fairness	Affected, Regulator	[8,12–14,16,22,24,32,38,40,44,51,53,55,56,58,60,63,64,67,69,70,78,81,86,87,96–100]			[29,43,68]
Informed Consent	Enable humans to give their informed consent concerning a system's decisions	Affected, Regulator	[15,56,57]			
Legal Compliance	Assess and increase the legal compliance of a system	Deployer	[5,7,12,15,16,21,22,24,26,27,29,30,34,35,44,47,55,56,58,62,63,66,68–70,74–76,81,87,91,98,100–102]			
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected, Regulator	[6,12,14,16,29,34,37,44,47,53,55,57,58,60,69,78]			
Performance	Assess and increase the performance of a system	Developer	[15,26,32,33,36,38,40,43,48,50,51,66,72,74,91,96,100,102]	[34,103]	[49]	
Privacy	Assess and increase a system's privacy practices	User	[14,16,78,98]			
Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	Regulator	[6,13,20,43,56,57,60,104]	[68]		
Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	Developer	[5,14,55,105]			
Safety	Assess and increase a system's safety	Deployer, User	[44,58,69,70,74,78,105]			
Satisfaction	Have satisfying systems	User	[5,13,15,24,28,33,36,39,43,46,78,79,94,102]			[31]
Science	Gain scientific insights from the system	User	[5,7,12,14,35,37,44,47,60,63,66,67,70,87,97,100,106]			
Security	Assess and increase a system's security	All	[55,57,78,100]			
Transferability	Make a system's learned model transferable to other contexts	Developer	[14,37,84]			
Transparency	Have transparent systems	Regulator	[13,16,25,26,29,33,34,37,39,44,46,49,55,57,62,63,67,74,75,78,80,92,94,107,108]	[28,31,43,50]		
Trust	Calibrate appropriate trust in the system	User, Deployer	[5–7,12,13,15,16,20,22,25–29,32–37,42,44–47,49,51,53,55–59,62,63,67,69–75,77–79,81,83,86–88,91,93–96,98,100,103,106–113]	[30,39]	[38,40,48]	[21,31,50,82,92,114]
Trustworthiness	Assess and increase the system's trustworthiness	Regulator	[7,12,33,38,41,60,69,74,77,91,94,104,105]			
Usability	Have usable systems	User	[13,22,28,33,36,43,57,62,63,78,104,108,110]			
Usefulness	Have useful systems	User	[13,34,36,43,63,77]	[31,45]		
Verification	Be able to evaluate whether the system does what it is supposed to do	Developer	[7,14,16,20,32,33,43,44,55,56,62,66,80]			

We classify the sources into those that provide no empirical investigation of their claims, those that show empirical evidence (e.g., an explainability approach affected a desideratum's satisfaction), those that provide mixed empirical evidence (e.g., for some explainability approaches there are positive effects on a given desideratum's satisfaction, whereas for others there are no effects), and those that present inconclusive empirical evidence (e.g., the effect of an explainability approach on a desideratum's satisfaction was not significant).

developers to recognize and correct such mistakes. Accordingly, there are cases where XAI contributes to determine whether a system works as intended and, thus, explainability approaches can support verification of the system.

Another important desideratum for developers is *performance*. There are many ways in which a system can achieve a better performance. For example, the predictive accuracy of an ML algorithm can be seen as a performance measure. Although there are some claims that explainability and accuracy are difficult to combine [40,44], there is also the opposite view, which sees XAI as a way to actually make systems more accurate and, in particular, to help developers estimate system accuracy [70,73]. By means of getting information of what led to a system's outcomes, developers can detect underrepresented or erroneous training data and, thus, fine-tune the learning process to achieve higher accuracy. Another way in which performance can be understood is user-system interaction. The better users can interact with a system, the better they, the system, and the combination of user and system perform. To this end, insights about a system, its functioning, and its outputs are a fruitful way to improve user-system performance [14,17,48].

*Affected Parties.* The influence of artificial systems is constantly growing and decisions about people are increasingly automated – often without their knowing. Affected parties are such (groups of) people in the scope of a system's impact. They are stakeholders, as for them much hinges on the decision of an artificial system. Patients, job or loan applicants, or defendants at court are typical examples of this class.

Crucial desiderata of affected parties are *fairness* [7,9,11,122] and *morality/ethics* [6,11,127]. These desiderata are closely related. If a system is fair, for instance, the influence of protected attributes (e.g., gender or ethnicity) is adequately limited or controlled in the systems' decision-making processes. In the case of ethical systems, their decision-making processes rely on morally permissible considerations (e.g., according to certain moral theories, an autonomous car in a dilemma situation should never let affected parties' age contribute to its decision-making process, see [128]).

Considerations of fairness and ethics have evolved because there is an increasing number of affected parties. This can lead to discrimination of individuals (e.g., concerning the distribution of jobs, loans, or healthcare), not on the basis of their own actions or characteristics but on the basis of actions or characteristics of social groups to which they belong (e.g., women, ethnic minorities, older people) [54]. One hope of establishing automated decision-processes was to make decisions less prone to human bias [129]. However, it is commonly acknowledged that artificial systems can reproduce and, in this process, even intensify human biases (see, e.g., [54] and [130]). To counteract biases, it is, therefore, crucial to enable their detection. Explainability approaches may aid in this regard by providing means to track down factors that may have contributed to unfair and unethical decision-making processes and either to eliminate such factors, to mitigate them, or at least to be aware of them.

*Deployers.* People who decide where to employ certain systems (e.g., a hospital manager decides to implement a diagnosis system in her hospital) are deployers. We count them as another class of stakeholders because their decisions influence many other classes of stakeholders. For example, users have to work with the deployed systems and, consequently, new people fall inside of the range of affected parties.

Deployers want the systems they bring into use to be accepted [5,11,96,131]. In the eyes of deployers, the worst case in terms of acceptance is that users reject appropriately working systems so that the systems will end up never being used [132]. Therefore, low acceptance undermines what deployers intend to achieve when providing systems to users. Previous research claims that explainability approaches can aid in this case by providing people with more insights into systems, which can improve their acceptance [11,115,131].

Another desideratum of deployers is the system's *legal compliance*. As deployers bear a certain degree of responsibility for systems they bring into use, they have to ensure that these systems comply with legislation. Non-discrimination and safety of a system are two important factors for its legal compliance. Explainability approaches promise to enable deployers and other stakeholders to check whether the system is indeed safe and non-discriminatory. Moreover, the European General Data Protection Regulation (GDPR) and the often discussed *Right to Explanation* [101] (arguably) explicitly require explanations.

*Regulators.* Finally, there are regulators stipulating legal and ethical norms for the general use, deployment, and development of systems. This class of stakeholders occupies a somewhat extraordinary role, since they have a 'watchdog' function not only with regard to systems, but to the whole interaction process of systems and the other stakeholder classes. This class consists of ethicists, lawyers, and politicians, who must have the know-how to assess, control, and regulate the whole process of using artificial systems.

Regulators call, for instance, for *trustworthy* systems [5,6,10,11,29,96,115,131]. However, the concept of trustworthiness is still only vaguely defined [133]. For example, the High Level Expert group on Artificial Intelligence (HLEGAI) initiated by the European Commission does not provide a common definition for trustworthiness, but it only proposes that trustworthy systems have three properties: they are lawful, ethical, and robust [10]. Without examining trustworthiness more closely, the HLEGAI emphasizes the significance of trustworthy artificial systems by stating that the trustworthiness of systems is imperative for the realization of potentially vast social and economic benefits. Regulators such as the EU, as well as previous research on artificial intelligence that calls for trustworthy systems (e.g., as described in [53]), agree that explainability approaches are one central way to facilitate the trustworthiness of systems [10,53].

*Accountability* is another important desideratum of regulators [20,56]. Accountability is about being able to identify who is blamable or culpable for a mistake. With increasing use of artificial systems, accountability gaps might emerge [134,135]. For instance, when the use of an artificial system harms a person, it may not be clear who is accountable, as there are many parties that may have contributed to the harm. Opaque artificial systems only amplify this issue. For example, a person acting on the outputs of a system may not (be able to) know that this output was erroneous, so blaming her

for ensuing problems might inadequately ignore the contribution of the artificial systems. Overall, regulators want to avoid situations in which existing legislation is hard to apply or where no one is (or feels) accountable for a mistake. In such cases, explainability approaches may restore accountability by making errors and causes for unfavorable outcomes detectable and attributable to the involved parties.

### 2.3. Interdisciplinary potential

Artificial systems will continue to influence humans in every part of their lives, thus it is likely that new desiderata will emerge. Further desiderata might evolve from societal, legal, political, philosophical, or psychological needs regarding artificial systems (e.g., for competence, relatedness, or autonomy; [136]). For example, with artificial systems in healthcare [137] there is a pressing need for formulating relevant ethical and legal desiderata. In addition, it is also possible that explanatory information provided by artificial systems does not only aim to improve task achievement but also to entertain users [138].

We conducted a literature review to derive an overview of stakeholders' desiderata, but it will clearly be possible to extend our list in Table 1. In fact, further developing and refining this list of desiderata is an important point that reveals interdisciplinary potential. First, most of the sources referred to in Table 1 only claim that these desiderata are relevant for stakeholders. There needs to be a more thorough empirical investigation, probably done by interdisciplinary teams of psychologists, philosophers, and scholars from law to show the actual importance of these desiderata for certain stakeholder classes.

Furthermore, in our overview, the desiderata's denotations stem (in most cases) directly from the source papers. However, some of these desiderata are closely related and, especially given the interdisciplinary research contributing to XAI, it is plausible that different authors actually mean to refer to the same desideratum but give it a different term or use the same term to refer to different desiderata. Consistent terminology and conceptual clarity for the desiderata are pivotal and there is a need to explicate the various desiderata more precisely. Different disciplines like law, philosophy, and psychology need to come together to discuss their conceptions of various desiderata to agree on common definitions of these desiderata. Without this, insights from different disciplines regarding the respective desiderata (e.g., what kind of explanatory information is required to satisfy a given desideratum) might not be adequately integrated into a common stream of research.

Additionally, we need research that more explicitly analyzes society's stakeholder classes affected by artificial systems. For instance, collaborating with sociologists could offer a broader or more nuanced picture of the classes of stakeholders that have to be considered within the scope of XAI. In any case, in order to comprehensively address the stakeholders' desiderata, we need a more detailed understanding of stakeholder classes and sub-classes. For this, it is promising to consult disciplines outside of computer science focusing on society (i.e., sociology) as well as individual differences within groups of society (i.e., psychology).

Furthermore, researchers from different disciplines may be able to take the perspective of certain stakeholder classes. By doing so, they can help to refine the list of desiderata. For instance, computer scientists can take the perspective of developers. Psychologists can take the perspective of users and affected people. Management scholars could take the perspective of deployers. Philosophers, political scientists, as well as researchers from law can take the perspective of regulators.

Working together in interdisciplinary teams can, thus, contribute to a comprehensive consideration of important desiderata in a given context. However, comprehensiveness is just one side of the coin, the justification of desiderata is another. Concerning this justification, there are two main perspectives: one from ethics and one from jurisprudence. From an ethical perspective, we can judge whether a desideratum is compatible with some, many, or even all established moral theories. Similarly, legislation can be consulted to assess whether there are laws demanding (or prohibiting) to meet certain desiderata. When engaging in thorough moral and legal justification, we might conclude that there will be desiderata that are not justifiable. In high-stakes decisions, for instance, each individual user might want systems to do what is best for her. In the case of autonomous cars, drivers will probably want a car to decide in a way that makes it more likely that they will survive if the car faces an imminent accident [139]. There might be cases where such a decision is neither morally nor legally justifiable. In a less drastic example, users may ask for an explanation of why they received a low score on a personnel selection test. However, providing this explanation might render the given test obsolete because the explanation possibly enables participants to game the test [8]. We suggest that the given context, as well as moral and legal considerations are decisive factors when determining whether certain stakeholders' desiderata can be justified.

## 3. Desiderata satisfaction requires understanding

In the previous section, we have introduced our claim that the need for XAI arises from stakeholders' desiderata. More precisely, the need arises in cases where certain stakeholders' desiderata are not (sufficiently) satisfied [7,14,15,18,53]. For this reason, we have to take a look at what it means for a desideratum to be satisfied.

### 3.1. Facets of desiderata satisfaction

We propose that the satisfaction of each desideratum can take two facets. We call these facets *epistemic* and *substantial* desiderata satisfaction, respectively. On the one hand, stakeholders want systems to have certain properties that make them

actually fair, transparent, or usable. In line with this, a desideratum (e.g., fairness) is substantially satisfied if a system sufficiently possesses the corresponding properties. On the other hand, stakeholders want to know or be able to assess whether a system (substantially) satisfies a certain desideratum (i.e., whether the system has the required properties). So, the epistemic facet of the fairness desideratum is satisfied for a stakeholder, if she is in a position to assess or know whether and to what extent the system is fair. Naturally, for XAI the epistemic facet is the most important one, since explanatory information can contribute to the satisfaction of the epistemic facet of every desideratum, whereas this is not the case for the substantial facet.

As an example, take the desideratum of having usable systems. A successful explanation process as depicted in our model may enable users to recognize whether a system is usable, and, optimally, also increase the system's usability to a certain degree. In this case, the epistemic satisfaction consists in the stakeholders being able to check whether a system or its outputs are usable for the task at hand. To a lesser extent, however, an explanation process can also contribute to the substantial satisfaction of the desideratum, since it provides additional knowledge about the system that makes it more usable for the stakeholder. For larger deficits in usability to be addressed, however, explanatory information might not directly help; for this, the entire artificial system may need to be redesigned.

Depending on the desideratum, the two facets are correlated to a certain degree (possibly even completely, when satisfying the epistemic facet also completely satisfies the substantial facet). To illustrate, consider the desideratum of retaining user autonomy in human-in-the-loop scenarios. Let us assume that an explanation process has helped to satisfy the epistemic facet of this desideratum to a certain degree, as it has enabled the user to assess the extent to which she can retain her autonomy in making decisions based on the recommendations of the system. Additionally, the more understanding a user has about a system's output, the more autonomous she can decide based on it. Thus, the explanation process has helped to satisfy the epistemic and the substantial facet of this desideratum. Hence, in this case, the substantial and the epistemic facet of desiderata satisfaction are highly correlated.

Now, consider the desideratum that systems adhere to certain ethical principles. When having sufficient information about a system, regulators can evaluate whether this system complies with ethical standards. Again, the explanation process serves to satisfy the epistemic facet of this desideratum. However, this does not directly make the system's processes and outputs more likely to comply with ethical standards. Consequently, explanation processes can at most indirectly satisfy the substantial facet of this desideratum: based on the understanding obtained by the explanation process, faults can be identified and steps to improve systems regarding their ethical properties can be initiated. In this case, the epistemic and the substantial facet of desiderata satisfaction are only loosely correlated.

On the one hand, the distinction of these two facets shows that explanation processes can contribute to the satisfaction of all epistemic facets of desiderata concerning artificial systems. On the other hand, it shows that an explanation process alone does sometimes not suffice to satisfy the substantial facet of desiderata concerning artificial systems. In many cases, however, the epistemic satisfaction enables the substantial one. This means that even if a better understanding of the systems triggered by explanatory information does not always directly lead to the substantial satisfaction of the desiderata, it can form the necessary basis for achieving it. As the epistemic satisfaction of a desideratum is closely linked to a better understanding of a system, understanding is the pivotal point for all endeavors of satisfying desiderata.

### 3.2. Understanding

Throughout the history of XAI research, authors have highlighted the central importance of understanding in XAI (e.g., [2, 20, 115, 140–142]). The overall goal of XAI is to advance human understanding of artificial systems in order to satisfy a given desideratum. There is an ongoing debate in the philosophical literature about what constitutes understanding [143–145], and a comprehensive review of this concept is beyond the scope of the current paper (see, for instance, [144] for a review on understanding, [146, 147] for papers on the concept of understanding, [138, 148, 149] for the relation between explanations and understanding, or [150] for the related topic of cognitive processes in knowledge acquisition; furthermore, see [151] for a broad overview on the theoretical basics of understanding relevant for XAI research). Some aspects of understanding, however, are typically agreed upon: There are different *depths* and *breadths* of understanding (in the following, we will use the term *degree of understanding* to address depth and breadth of understanding, similar to [144, 152]), and there are different *kinds* of understanding [20, 146].

For the evaluation of explainability approaches it will, thus, be crucial to determine stakeholders' understanding of artificial systems. Examining understanding of software has a long history in human-computer interaction and education [108, 153]. Typically, when referring to the understanding of an artificial system, these disciplines also refer to users' *mental models* of systems. A mental model of a system can be understood as a mental representation of this system and its functioning. According to Rouse et al., mental models allow humans 'to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future states' [154]. In other words, mental models allow humans to mentally simulate aspects of a system, for instance, in order to understand the causes of its decision-making [17].

Translated to the case of artificial systems, this means that we could examine understanding by investigating stakeholders' mental models of a system [17, 108]: how well does a person's mental model mirror the actual system? Are there gaps in the current understanding of a system's functioning? Are people overconfident that they understand how a system works (when they actually only have an illusion of understanding) [155]? Are there learned misconceptions about systems

and their outputs that need to be revised [156]? For example, it is possible to investigate a stakeholder's mental model through think-aloud techniques where stakeholders are tasked to describe systems and their inner workings, or by letting stakeholders draw their mental model of a given system (for an overview of methods to elicit mental models, see [17]). Similarly, it is possible to measure understanding by capturing what humans' mental models enable them to do in relation to a system and its outputs. For example, Kulesza et al. [157] used an *explanation task* to assess whether participants understood what kinds of information are used to predict outcomes. Other studies used *prediction tasks* to assess, for instance, whether participants can anticipate which predictive model would produce better outcomes [115], whether participants can predict what outcome a predictive model would produce for a person with a given profile [21], or whether participants can foresee the influence of a given feature on an outcome [21]. Another possibility would be to use *manipulation tasks* in order to assess whether people understood what kind of information might add to the predictive accuracy of a model (e.g., like in [115]). Further possible tasks might be *perception tasks* (e.g., naming of recognized characteristics of a model) or *imagination tasks* (e.g., estimating what a model would predict for a given input), all of which would reflect different degrees of understanding. Furthermore, all of these tasks can reveal misconceptions of a system's functioning or knowledge gaps that need to be adjusted or filled with additional or alternative explanatory information. Additionally, what all of these ways to capture stakeholder understanding have in common is that they might help us to examine whether a given desideratum has the potential to be satisfied. For example, if a developer has understood a system in a way that she can imagine situations under which a system might fail, her ability to make the system more robust most likely increases.

Furthermore, there is an initial degree to which stakeholders understand artificial systems. Specifically, stakeholders without any prior experience with a given system will likely start with a degree of understanding that corresponds to their (background) knowledge of artificial systems or arose from initial instructions they have received regarding the system [17, 158]. Thus, they will have an incomplete or even faulty mental model of the given system [20]. For instance, a stakeholder might know (or might be informed) that machine-learning based systems are usually trained on historical data in order to predict new data. This degree of understanding can, then, be augmented (e.g., with explanatory information generated from explainability approaches). With a higher degree of understanding (and, consequently, a more detailed and accurate mental model of a system), a stakeholder might understand what kind of training data underlie a given system, what kind of algorithm is used for a given system, or what kind of output data a system produces [17, 63, 159]. Thus, with increasing degrees of understanding, stakeholders will be able to assess whether a given system has desired characteristics and adequate processes, or produces expected outcomes. In other words, an increasing degree of understanding will satisfy the epistemic facet of more and more desiderata. For the satisfaction of the desideratum's substantial facet, however, the opposite might sometimes be the case, as we will discuss in the next section.

### 3.3. The relation between understanding and desiderata satisfaction

In certain cases, a stakeholder's degree of understanding and the extent of desiderata satisfaction are positively correlated. For instance, if the desideratum is to retain autonomy in interaction with a system, usually a higher degree of understanding satisfies the epistemic and substantial facets of this desideratum to a greater extent. However, there are also more complex cases. Assume that the desideratum is to trust a certain system. Acquiring a higher degree of understanding will increase a stakeholder's epistemic satisfaction of this desideratum (i.e. she can better assess whether and to what extent to trust the system), but the substantial facet (i.e., her actual trust) can be influenced in a negative way. When a stakeholder still possesses a low degree of understanding, she is likely to be unaware of problematic features a system has in certain contexts (e.g., in complex environments) or with certain kinds of input data (e.g., noisy inputs). So, with a low degree of understanding, a stakeholder is likely to trust a system (although inadequately) [29, 38]. In contrast, with a higher degree of understanding, the stakeholder is able to recognize or even explain the conditions under which a system will tend to fail. Therefore, she is more aware of the system's problematic features, and this may, consequently, decrease her trust in it [21, 38].

Additionally, it can happen that understanding contributes to the satisfaction of a single desideratum of a stakeholder to a greater extent, while the satisfaction of other desiderata for the same stakeholder suffers [63]. To illustrate, take the trade-off between transparency and non-gameability of systems. Deployers of systems want to comply with legislation and, consequently, want their systems to be transparent. Understanding is a necessary condition for perceived transparency. However, making systems more transparent can diminish another desideratum of deployers, the systems' non-gameability [8]. In other words, it should not be possible for particular users to manipulate a system in such a way that they can systematically evoke beneficial outputs for them. However, more transparency caused by a higher degree of understanding may enable some people to exploit the system [63]. In a personnel selection test, for example, a better understanding of the selection system may enable participants to game the test, preventing its proper use (i.e., selecting suitable applicants).

This points to another potential trade-off to be considered regarding the relation between an advanced understanding and desiderata satisfaction. Felzmann and colleagues [15] argue that different stakeholders might hold different expectations regarding the extent to which a single desideratum has to be satisfied. It is furthermore possible that, while desiderata of one stakeholder are influenced positively by an increase in understanding, desiderata of other stakeholders suffer. An example of such a case was described by Langer et al. [160]. They provided additional information accompanying an automated

personnel selection system for their participants, which resulted in more perceived transparency, but at the same time reduced acceptance of the system (for a similar finding see [161]). In such cases, it can happen that the two desiderata of transparency and acceptance arise from different perspectives and characteristics of stakeholders. For instance, legislation (i.e., a regulator) might call for transparency of systems, whereas a company using a system (i.e., a deployer) desires the system to be accepted. Explaining the system will, then, lead to the satisfaction of the legal desideratum, but at the price of impairing the company's desideratum.<sup>4</sup>

These examples indicate that the degree and kind of understanding of artificial systems which explainability approaches should evoke may depend on trade-offs between a variety of desiderata from a variety of stakeholders. Consequently, the development, implementation, and use of explainability approaches should go hand in hand with a case-by-case evaluation of the relevant stakeholders' desiderata. While estimating the effects of explainability approaches, it is central to investigate the perspective of not only one, but all stakeholders who potentially have a variety of (conflicting) desiderata with regard to a given system.

### 3.4. Factors influencing the relation between understanding and desiderata satisfaction

Characteristics of the context in which artificial systems operate moderate the relation between understanding and desiderata satisfaction and may affect the degree and kind of understanding necessary to satisfy a given desideratum. There is no agreed-upon definition of the term 'context' [162] and a deeper dive into the discussion about the term goes beyond what we can achieve in this paper (for discussions on this topic, see [162,163]). Following Dourish [163], we hold that the context is set by a given situation, in the interaction between a stakeholder, an artificial system, a given activity or task, and an environment. This makes it impossible to anticipate all contextual influences that will affect the process of how explainability approaches aim to satisfy stakeholders' desiderata without knowing the concrete situation. Nevertheless, it is crucial to consider and anticipate these contextual influences when evaluating and developing explainability approaches.

For instance, what is at stake in a given situation can affect the relation between understanding and stakeholders' desiderata. That is, whether a context is a high or low stakes scenario may determine the degree of understanding necessary to satisfy a given desideratum. Research indicates that certain situations tend to require a greater understanding of an event than other situations. Specifically, situations where instrumental, relational, moral, or legal values are at stake might be more likely to require extensive understanding [164–166]. Instrumental values are at stake when there are (personal, economical etc.) benefits or losses to expect in a specific situation (e.g., when an artificial system handles financial transactions). Relational values are at stake when important interpersonal relationships might be affected by an event (e.g., when artificial systems are used for employee layoff). Moral values are at stake when moral rights might be violated (e.g., when using artificial systems for sentencing in court). Finally, legal values are at stake when legal rights might be violated (e.g., when an artificial system's outputs conflict with the right of non-discrimination). Depending on the concrete situation (e.g., being in an autonomous car), instrumental, relational, moral, and legal values determine the stakes of a situation. Additionally, identifying which of these values stakeholders regard as relevant or which values are indeed relevant in which situations may allow for drawing inferences about whether a situation is (considered) high or low stakes. Consequently, these values serve as an orientation for when stakeholders are more likely to demand higher degrees of understanding. Specifically, this means that for the satisfaction of a given desideratum in a low stakes scenario, a lower degree of understanding might be sufficient compared to high stakes scenarios.

Furthermore, contexts involving artificial systems may differ in their 'outcome favorability' [167,168]. In the case of an unfavorable outcome, people are more likely to call for additional information in order to understand the reasons for the outcome, and so to be able to better assess and control further similar outcomes [36,167]. On the one hand, it may be that increasing the extent of understanding in the case of favorable outcomes has little potential to positively affect a desideratum's satisfaction. On the other hand, a better understanding can be central to positively affecting a desideratum's satisfaction in the case of an unfavorable outcome. For instance, perceived fairness is a central desideratum for many decision situations [169]. While understanding may have negligible effects on perceived fairness under favorable outcomes, it can improve perceived fairness under unfavorable outcomes [170]. Supporting this claim, Kizilcec [38] found that available explanatory information only affected perceived fairness when people's expectations concerning an outcome were violated. There may, however, be certain conditions under which a better understanding (negatively) affects perceived fairness even when people experience favorable outcomes [166].

The stakes of a situation and outcome favorability are just two of many possible contextual factors that may influence the relation between understanding and desiderata satisfaction. Further candidates are the application context (e.g., at home, at work), time constraints [171], and social circumstances (e.g., whether there are other people present in a given situation; [163]).

---

<sup>4</sup> Note that this example also describes a situation where two desiderata of the company are in conflict: user acceptance and adhering to legislation.

### 3.5. Interdisciplinary potential

The interdisciplinary potential we see in the relation between understanding and desiderata satisfaction is described in research questions such as: (a) How does stakeholders' initial degree of understanding or prior knowledge of artificial systems relate to their desiderata satisfaction? (b) What are the trade-offs between desiderata of a single stakeholder and/or desiderata of multiple stakeholders and what are the implications of understanding regarding these trade-offs? (c) How does the degree and kind of understanding relate to the satisfaction of desiderata? (d) How do task or context affect the relation between understanding and desiderata satisfaction in a given situation?

Scholars from educational sciences could collaborate with computer scientists in order to investigate how to design adequate instructions to achieve a proper basic understanding or background knowledge of artificial systems that can serve as a general basis to partially satisfy a large variety of desiderata. Furthermore, it is necessary to involve psychologists in order to experimentally examine the relation between understanding and desiderata satisfaction as well as influences affecting this relation. In this regard, it will be central to determine what it means for a certain desideratum to be satisfied. This requires finding an adequate way of measuring the satisfaction of this desideratum (e.g., using self-report measures, expert interviews, legal analyses), as well as understanding the requirements for it to be satisfied (e.g., defining minimum legal standards, enabling stakeholders to perform a specific task successfully). In practice, this involves having an elaborated research design, clear conceptual definitions, appropriate operationalization and measurement methods, and fitting research disciplines for iterative (empirical) research.

Finally, scholars from requirement engineering may help to understand relationships between several desiderata. Based on their results, scholars from law and from philosophy can help to determine which trade-offs are morally and/or legally justified. In general, an interdisciplinary collaboration can contribute to being aware of potential relationships and trade-offs between certain desiderata.

## 4. Understanding requires explanatory information

Providing explanatory information of a given phenomenon is the default procedure to facilitate its understanding [147,172,173]. Explanatory information helps humans to navigate complex environments by facilitating better understanding, predictions, and control of situations [138,174]. Such information narrows down possible reasons for events, decreases uncertainty, corrects misconceptions, facilitates generalization and reasoning, and enables a person to draw causal conclusions [138,155,174,175].

In the context of XAI, explanatory information puts stakeholders in a position to grasp generalizable patterns underlying the production of a system's outcomes (e.g., LIME [115]). These patterns allow for drawing inferences about (potentially causal) connections between the system's inputs and outputs, or for narrowing down the possible ways in which the system might have failed. Additionally, these patterns may help stakeholders to tell apart correct but unexpected behavior from malfunctions in order to debug the system. In general, all of this can reduce people's uncertainty concerning a system (e.g., uncertainty concerning how to behave towards it, how to react to it, what to think of it, whether to recommend it, or whether to disseminate it; [63]). Overall, explanatory information should lead to a better understanding, which should, in turn, positively affect the satisfaction of stakeholders' desiderata. Depending on the explanatory information, different degrees and kinds of understanding can be acquired [20,152]. For this reason, we have to shed more light on what characteristics of explanatory information are.

### 4.1. Characteristics of explanatory information

Important characteristics of explanatory information concern its kind and its presentation format. There are various kinds of explanatory information: teleological (i.e., information that appeals to the function of the explanandum; [176]), nomological (i.e., information that refers to laws of nature; [177]), statistical relevance (i.e., information that is statistically relevant to the explanandum; [178,179]), contrastive (i.e., information that highlights why event P happened and not event Q; [6]), counterfactual (i.e., information that appeals to hypothetical cases in which things went differently; [180]), mechanistic (i.e., information that appeals to the mechanisms underlying a certain process; [181]), causal (i.e., information that appeals to the causes of an event; [182,183]), network (i.e., information that appeals to the topological properties of a network model describing a system; [184]), and many more.

Concerning the presentation format, there are also various possibilities. Roughly, we can distinguish between text-based and multimedia presentation [36]. A text-based presentation can be a natural language text, a rule extracted from a rule-based system, an execution trace, or simply the program's source code. A multimedia presentation can include graphics, visualizations, images, and animations. For instance, heat-maps of neural activity are a popular presentation format for explanatory information of neural networks [14].

Aside from the very general characteristics of kinds and presentation format, there are further characteristics of explanatory information that influence how and whether it evokes understanding: soundness (i.e., how accurate the information is), completeness, novelty (i.e., whether the information is new for the recipient), and complexity (e.g., depending on the number of features and on the interrelation between the features the information contains; [171]) are just some of the various examples of further characteristics of explanatory information (for more, see [63]).

The importance of acknowledging the characteristics of explanatory information is highlighted by research in cognitive and educational psychology that shows that effects of explanatory information can vary depending on their characteristics [138,150,166]. Take complexity as an example: studies show that people prefer simpler information (e.g., information mentioning fewer causes; [185]). Another example is the finding that explanatory information that aligns with a stakeholder's goals in a certain situation is preferred [186]. Vasilyeva et al. [186] showed that people evaluate teleological explanatory information compared to mechanistic explanatory information as more useful when asked to name the function of a phenomenon, and, conversely, perceive mechanistic explanatory information as more useful when asked to name the cause of an event.

#### 4.2. Factors influencing the relation between explanatory information and understanding

The relation between explanatory information and understanding is influenced by a variety of characteristics of the stakeholders, the context, and interactions between these characteristics [164,166,187]. Considering these influences is central to evaluate and develop explainability approaches.

*Characteristics of stakeholders.* Since every stakeholder possesses an individual degree of understanding of a system, an individual ability to understand, and an individual set of desiderata that are or need to be satisfied to a certain extent, the characteristics of stakeholders who receive explanatory information influence the relation between explanatory information and understanding [18]. Some of the characteristics that most obviously influence this relation are the stakeholders' background knowledge, beliefs, learning capacities, and desiderata they have concerning respective systems [63,188–190]. For instance, explanatory information including technical details might increase an expert developer's degree of understanding while technical details can hamper understanding for novice developers or other (non-expert) stakeholders [175].

Furthermore, desiderata that are salient for a respective stakeholder can influence the relation between explanatory information and understanding. Specifically, as stakeholders engage with information in order to advance their understanding of artificial systems, their motivation and prior beliefs may affect how they interpret a given set of information [67]. For instance, if a stakeholder's primary desideratum is to ensure that a system provides fair outputs, they will scrutinize explanatory information for signs of bias that might lead to unfair outcomes. In contrast, if a stakeholder's primary desideratum is to improve a system's predictive accuracy, they might pay special attention to information providing insights on how to improve system performance. This indicates that, depending on the given desideratum, the same amount and kind of information can lead to different degrees and kinds of understanding. Consequently, it is important to provide the appropriate information for the given purpose. These assumptions are supported by research proposing that human reasoning processes can be strongly influenced by explanatory information. Such information has the potential to improve human decision-making but may also hamper it (e.g., when explanatory information attenuates human biases or when humans need to invest too much effort to use the information [36,67]).

The list of individual differences influencing the relation between explanatory information and understanding goes beyond the scope of this article and covers personality traits such as conscientiousness [191], need for cognition<sup>5</sup> [192,193], and need for closure [194,195], as well as people's preferences for detailed versus coarse explanations [196], or age differences between stakeholders [197].

*Characteristics of the context.* Time pressure can be a relevant contextual influence [63,171]. For a stakeholder under high time pressure, the same explanatory information may lead to less understanding as compared to situations where she is under low time pressure [198]. Workload is another contextual influence [199]. In situations of high perceived workload, the same explanatory information can lead to a different degree of understanding compared to low workload conditions. Similar things are true for situations where it is more likely that stakeholders will experience higher levels of stress [200,201] (e.g., high stake and high risk situations, multitasking environments; [171,202]).

In general, depending on the situation and task at hand, the effects of explanatory information on understanding may differ. Therefore, it is important to investigate the contextual conditions of a stakeholder's interaction with an artificial system in detail when theorizing about how explanatory information can best improve understanding. Given the impact of the context and given the fact that it might not be possible to anticipate all relevant contextual influences [163], it is especially important to assess the effects of explanatory information in laboratory and in field settings [9,63]. Although it has advantages to investigate the effects of, for instance, stress on the relation between explanatory information and understanding in controlled laboratory settings, such results may not translate to field settings. This finding is in line with calls for experiments involving human participants in proxy tasks in order to show that given explanatory information not only elicits understanding when simulating a context, but also under real world conditions (that is, to investigate whether given explanatory information is equally valuable in the wild) [9,63]. Although it will be impossible to fully anticipate how the context will alter the relation between explanatory information and understanding (e.g., because the interpretation of

<sup>5</sup> Need for cognition is a personality trait that distinguishes people who like to put effort into cognitive activities from those who prefer less cognitively demanding activities.

contextual influences depends on the relevant stakeholders [162]), at least it is crucial to be aware that contextual influences may be central for the success or failure of explanatory information.

*Interactions between characteristics of explanatory information, stakeholders, and the context.* Finally, characteristics of explanatory information, stakeholders, and the context can interact in ways that affect how stakeholders engage with explanatory information to advance their understanding of artificial systems. For instance, the level of detail of explanatory information can interact with the prior knowledge of a stakeholder. Whereas expert developers' understanding may benefit to a higher degree from explanatory information with much detail, this level of detail can have the contrary effect for novice developers [36,166,203,204]. At the same time, when novice users want to learn how to use a system, they might want detailed explanatory information whereas when expert users want to use the system for task fulfillment, every unnecessary piece of information could lead to the rejection of the system [36]. Further, if the context changes, the relations between explanatory information and understanding are prone to change as well. For instance, as soon as there is time pressure in the aforementioned situations, it is plausible that neither expert nor novice developers or users pragmatically benefit from too detailed explanatory information.

#### 4.3. Interdisciplinary potential

The following research questions reinforce calls for extensive validation and experimental studies investigating the effects of explainability approaches in different contexts and in relation to different stakeholders [9,63] (we refer readers to Sokol et al. [63] for a description of further important characteristics of explanations, stakeholders, and contexts that affect the relations between explanations and stakeholders' understanding): (a) How should we classify explanatory information? (b) How does different explanatory information lead to understanding? (c) How do different stakeholders engage with explanatory information? (d) How can we optimally evoke understanding through explanatory information? (e) How do stakeholder differences (e.g., background knowledge, personality characteristics) affect understanding? (f) How do contextual influences (e.g., different levels of risk, multi-tasking environments, time pressure) affect the relation between explanatory information and understanding? (g) How should explanatory information be designed? What should it include? What kind of presentation format (e.g., textual, graphical) is appropriate? How can stakeholders interactively engage with this information?

On the one hand, these research questions call for a unified classification of explanatory information. Often, computer scientists classify explanatory information based on its presentation format or based on the explainability approach it originated from (e.g., [205]). Philosophers and psychologists, however, usually classify explanatory information in terms of the kinds mentioned above (e.g., causal or nomological; [206]). In order to prevent the debate from drifting apart, philosophers, psychologists, and computer scientists should collaborate to find standardized ways to classify explanatory information. On the other hand, these research questions call for empirical evaluation of the value of different kinds of explanatory information for different stakeholders, under a variety of contexts and under the consideration of different desiderata. This may primarily call for empirically working psychologists, as they have the tools and expertise to design experimental studies, for philosophers to determine the qualities of good explanatory information, and for computer scientists to adjust the presentation of explanatory information as well as possible. Furthermore, cognitive scientists might need to examine what exactly understanding of artificial systems actually means and how to measure it (see [17] for ideas of how to capture human understanding of artificial systems).

However, deriving explanatory information of artificial systems is a task that in itself requires a lot of interdisciplinary research. The following section completes the specification of the concepts and relations in our model by describing that, in order to get the required explanatory information from the systems, there is the need for fitting explainability approaches.

### 5. Explanatory information requires explainability approaches

In order to provide explanatory information that facilitates understanding and, thus, affects the satisfaction of the desiderata of the different classes of stakeholders, XAI research has developed a wide variety of explainability approaches. These approaches encompass methods, procedures, and strategies that provide explanatory information to help stakeholders better understand artificial systems, their inner workings, or their outputs. A specific explainability approach is characterized by all steps and efforts that are undertaken to extract explanatory information from a system and to adequately provide it to the stakeholders in a given context. Explainability approaches can take many guises and the literature commonly distinguishes two families of approaches (e.g., [12,14,63,131,205,207]): *ante-hoc* and *post-hoc* approaches.

#### 5.1. Families of explainability approaches

Ante-hoc approaches aim at designing systems that are inherently transparent and explainable. They rely on systems being constructed on the basis of models that do not require additional procedures to extract meaningful information about their inner workings or their outputs. For example, decision-trees, rule-based models, and linear approximations are commonly seen as inherently explainable (given they have a limited size) [7,131]. A human can, in principle, directly extract information from these models in order to enhance her understanding of how the system works or of how the system arrived at a particular output. Unfortunately, this way of deriving explanatory information from transparent models

might only be useful for stakeholders with a certain expertise [16,208]. For this reason, ante-hoc approaches make systems directly explainable only for developers and members of other stakeholder classes that possess enough expertise about artificial systems. Furthermore, ante-hoc explainability can also lead to a loss of predictive power [14,15] and not all systems can be designed inherently explainable.

Post-hoc approaches try to circumvent the aforementioned shortcomings. Such approaches are, in principle, applicable to all kinds of models. The difference to ante-hoc approaches is that post-hoc approaches do not aim at the design-process of a particular system, but at procedures and methods that allow for extracting explanatory information from a system's underlying model, which is usually not inherently transparent or explainable in the first place [11,63,131,207]. Post-hoc approaches are, for example, based on input-output analyses or the approximation of opaque models by models that are inherently explainable.

In many cases, however, post-hoc approaches are restricted with respect to how they present explanatory information. That is, given a specific model or one of its outputs, the information an approach will provide on repeated usage (and the format in which the approach provides the information) will be similar. Hence, for post-hoc approaches the same holds as for ante-hoc approaches: it is not guaranteed that all stakeholders are able to understand the provided information in the given format [16,208]. So, the explanatory information accessible from both, post-hoc and ante-hoc approaches is often only interpretable for developers or other expert stakeholders. This means that this information does not directly facilitate understanding for non-experts [16].

One solution to this is to combine several explainability approaches in order to cover a broad range of different information and presentation modes. Another solution that has received increasing attention is to have recourse on interactive explainability approaches [12,67]. Interactive approaches are based on the idea that the user or some other stakeholder is provided with more in-depth information concerning a system if the information she initially received does not suffice. Based on her needs, the person interacting with the system can call for information about specific aspects of a decision or request a presentation in a different format. To date, however, approaches that are fully interactive remain rare [12,205].

A third solution is to have a human facilitator (e.g., an expert stakeholder) explain a system to other stakeholders. For instance, when regulators want to satisfy their desiderata concerning artificial systems, there will be cases where they do not directly interact with an artificial system. Instead, a human facilitator (e.g., an expert user or a developer) will do so and derive suitable explanatory information for the regulator. In a sense, this process introduces a desiderata hierarchy based on the stakeholders into Fig. 2. In other words, one desideratum might have to be satisfied for one stakeholder class before some other desideratum (for another stakeholder class) can be satisfied. For example, the facilitator's desideratum (in this case: to be able to explain a system to a regulator), must be satisfied first as a precondition for the satisfaction of a regulator's desideratum (e.g., to be able to assess the fairness of a system). Thus, we have the complete processes in Fig. 1 nested within the explainability approach. Having a human facilitator (e.g., a developer) deriving explanatory information (assisted by an explainability approach) for another stakeholder (e.g., a regulator) can be considered a hybrid human-system approach to explainability.

## 5.2. Factors influencing the relation between explanatory information and explainability approaches

There are further characteristics of explainability approaches that are worth mentioning, since they are likely to influence the provided kind of explanatory information or its presentation format. First, it is important to distinguish post-hoc approaches that work regardless of the underlying model type (so-called *model-agnostic* approaches) from ones that only work for specific (types of) models (so-called *model-specific* approaches). Model-agnostic approaches aim to deliver explanatory information about a system solely by observing input/output pairs [63,115,131,207]. Model-specific approaches do so while also factoring in specific features of the model at hand (e.g., by creating prototype vectors in a support vector machine) [63,131,207]. Model-agnostic approaches have the advantage of working for all types of models, but have the drawback that they tend to be less efficient, less accurate as well as less explanatory powerful (i.e., the explanatory information's level of detail is lower with regard to individual phenomena) than the former.

Second, previous research distinguished the scope of an explainability approach. Some approaches provide information about only single predictions of the model. The scope of these approaches is *local* [63,115,131,207]. Often, they offer visualized prototype outcome examples (e.g., [209,210]). The more general type of approaches has a *global* scope [63,131,207]. These approaches are designed to uncover the overall decision processes in the model. Here, the usual way to provide this information is by approximating complex models with simpler ones that are inherently explainable.

Depending on the explainability approach, the explanatory information provided will differ. Global explainability approaches, for instance, are likely to produce more complex information that requires more background knowledge by stakeholders to be understood. Local explainability approaches, on the other hand, only show a limited picture of a system's inner working and may not be representative of its overall decision-making processes. Based on these differences we can conclude that certain explainability approaches are more suitable for the satisfaction of given desiderata of specific stakeholder classes than others.

To elaborate, take users who want to calibrate their trust in a system. They will need a different kind of information compared to users who want to have usable systems. In the first case, the explainability approach will likely need to extract information about the robustness of a system, about conditions under which outputs of the system are trustworthy and situations where users have to be aware that outputs might be misleading. In the case of usable systems, users might

want information that directly matches their specific goals in a given task. Another example: if a desideratum of users is learning how to use a system, they may need a different kind of explanatory information compared to when they simply want to fulfill tasks with the help of an artificial system [211]. This is because users who want to learn how to use a system need more details whereas users who want to fulfill tasks need directly useful information in order not to reduce their productivity through overly detailed information. Differences in the information needed can also be due to differences in the perspective of stakeholders. Take the desideratum of fairness. Affected parties will more likely focus on aspects of individual fairness which may call for explainability approaches that facilitate local explainability. Regulators, however, may focus on more general notions of fairness calling for explainability approaches facilitating global explainability.

### 5.3. Interdisciplinary potential

The design of explainability approaches and the goal of providing adequate explanatory information with the potential to affect stakeholders' desiderata, again, hold untapped interdisciplinary potential, with research questions such as: (a) How can we pinpoint what explanatory information an explainability approach should provide in which case? (b) How can we design interactive explainability approaches? (c) How can explainability approaches involving human facilitators be optimally designed? (d) How can we guarantee that an explainability approach provides the required explanatory information? (e) How should stakeholders' desiderata and their degrees of understanding be taken into account when generating explanatory information or when developing new explainability approaches?

For all of these research questions, we see a potential for collaboration between computer scientists, philosophers, psychologists, and cognitive scientists. For instance, philosophers and psychologists have to determine which information is needed to assess whether a system is fair, whereas computer scientists develop explainability approaches that provide this information, are aware of trade-offs between different approaches as well as of technical constraints. Furthermore, investigating how the examination of stakeholders and their desiderata narrows down the options of possibly successful explainability approaches might be a fruitful area for future research.

A more thorough discussion about which desiderata of what stakeholder class call for what kind of explainability approach goes far beyond what a single paper can achieve. However, in Section 6, we will show that our model could inspire work that is necessary to evaluate the usefulness of an explainability approach in relation to a given desideratum. Furthermore, we will show that the model supports the development of new explainability approaches to satisfy a given set of desiderata. Thus, the next section aims to show how our model can lead to actionable insights for XAI research by analyzing the concepts and relations that we propose in Fig. 1. By means of hypothetical application scenarios, we want to stimulate ideas about specific applications of our model.

## 6. Bringing it all together: hypothetical application scenarios

Following, we will present how the previous sections come together within hypothetical application scenarios. These scenarios highlight the importance of different stakeholder classes and their desiderata. Furthermore, these scenarios emphasize that understanding affects the satisfaction of these desiderata, that explanatory information provided by explainability approaches facilitate understanding, and that analyzing and investigating these concepts and their relations is central for the aims of XAI as well as for the development of explainability approaches that can successfully satisfy stakeholder desiderata.

The main application of our model is derived from the idea that if an explanation process does not change a certain desideratum's extent of satisfaction, the corresponding explainability approach might not be a suitable means for satisfying the desideratum in the given context. Such a discovery (e.g., resulting from stakeholder feedback or from empirical investigation) can provide feedback regarding which explainability approaches and what kinds of explanatory information work for which desiderata in which contexts. Thus, this feedback can serve as an input for the improvement of explanation processes and, consequently, helps to evaluate, adjust, choose, and develop explainability approaches for a given purpose and context.

We propose that each step in our model (Fig. 1: Explainability Approaches → Explanatory Information → Understanding → Desiderata Satisfaction) allows for drawing inferences about the explanation process involving explainability approach(es), kind(s) of explanatory information, and stakeholder understanding. The following questions, which arise at different points in our model, are of particular interest:

- Who are the relevant stakeholders and what are their specific characteristics? Which are the relevant desiderata in a specific context and are they satisfied?
- Have the stakeholders acquired a sufficient degree and the right kind of understanding that allows for assessing whether given desiderata are satisfied and to facilitate their satisfaction?
- Does the provided explanatory information and its format of presentation facilitate stakeholders' understanding in a given context and in consideration of the stakeholder characteristics?
- Is the explainability approach able to provide the right kind and amount of explanatory information in the right presentation format?
- Are there contextual influences hindering or promoting the satisfaction of desiderata through the explanation process?

Investigating these questions requires empirical research, hypothesis testing and interdisciplinary cooperation, but should, eventually, aid in evaluating, adjusting, choosing, and developing explainability approach(es) and finding explanatory information in order to adequately satisfy stakeholders' desiderata.

With this in mind, we believe that our model is useful for several important application scenarios. First, our model is useful for choosing adequate explainability approaches for novel application contexts of artificial systems and for guiding the development of new explainability approaches. Specifically, our model can be used to inform projects on how to develop explainability approaches to satisfy certain desiderata for a given class of stakeholders. Second, our model is useful for evaluating why and at which stage an explainability process failed to contribute to satisfying the relevant desiderata. Let us assume that the use of an explainability approach does not lead to the satisfaction of a certain desideratum. Why is this the case? Is the explainability approach not suitable for the satisfaction of the desideratum (e.g., the explainability approach provides the false kind of explanatory information) and should be replaced by another approach? Or does the error lie somewhere else in the explanation process? In some cases, we may be able to adjust the explainability approach appropriately to achieve its intended purpose.

### 6.1. General application scenarios

For structured attempts to evaluate, adjust, choose, or develop explainability approaches for a given context we roughly distinguish two scenarios, which we call the *evaluation* and the *discovery* scenario. In the evaluation scenario, we want to investigate whether the use of a specific explainability approach was adequate, and if not, what is needed to fix its shortcomings. In the discovery scenario, we want to find an adequate explainability approach to satisfy stakeholders' desiderata. This can take one of two forms: choosing among existing approaches or, if no adequate approach is available, developing a new one.

*Stakeholder and Desiderata.* Both evaluation and discovery scenarios start with examining the stakeholders and clarifying their desiderata in the given context. In the discovery scenario, we have to examine what the relevant classes of stakeholders are and which desiderata they have concerning the application of a system under consideration. In the evaluation scenario, we have to check – even in scenarios where already identified desiderata are satisfied – whether the explanation process fits all relevant classes of stakeholders and all of their desiderata (and ensure that we did not overlook important stakeholders or desiderata).

For this, we need input from a wide variety of disciplines including but not limited to scholars from law, sociology, psychology, philosophy, and computer science. Such a combination of expertise and perspectives helps to identify relevant stakeholder classes and list their desiderata pertaining to a given context. Defining these desiderata falls within the expertise of philosophers. The elaboration of the desiderata's relevant moral and legal aspects is a task for ethicists and scholars from law. Furthermore, assessing contextual peculiarities and stakeholder characteristics relevant in a given context will require psychologists (on the individual level), sociologist (on the group level), as well as domain experts such as judges or personnel managers (for particular application scenarios).

*Desiderata Satisfaction.* For the discovery scenario, the next step is to determine what it means that stakeholders' desiderata are satisfied in a specific context: to make estimates, to provide guidelines, and to create measures for when the identified desiderata are satisfied. For the evaluation scenario, we have to check whether the relevant desiderata are satisfied or not. If they are, the explanation process was successful and the chosen explainability approach appropriate. If they are not substantially satisfied, there are two possible cases. First, the necessary understanding was not acquired and, thus, the desideratum is also not epistemically satisfied. Then, an improvement of the stakeholders' understanding is required. Second, an adequate degree and kind of understanding is reached and, thus, the epistemic facet of the desideratum is satisfied. In this case, we may conclude that the regarded desideratum is not directly substantially satisfiable by means of explainability approaches. Regardless of the case, at this point we have to move on to investigate stakeholders' understanding.

Determining conditions for desiderata satisfaction will be an interdisciplinary task for psychologists, philosophers, and scholars from law. Furthermore, computer scientists and domain experts can give practical input on satisfaction conditions for the desiderata in their specific domain. Making the satisfaction of these desiderata measurable and examining their extent of satisfaction will be a job for psychologists who develop measures or tasks that help to assess the extent of desiderata satisfaction. Additionally, it can be a task for scholars from law or philosophers to provide clear guidelines for when a desideratum is satisfied.

*Stakeholders' Understanding.* The discovery scenario continues by investigating and defining requirements for the stakeholders' understanding needed to satisfy the desiderata under consideration. Specifically, we have to determine the appropriate degree and kind of understanding concerning the system and its output that promise to enable the epistemic facet of desiderata satisfaction. At the same time, this means that we need to assess (e.g., in studies using tasks to measure stakeholders' mental models of a system, or in studies using tasks to reveal whether stakeholders were able to, for instance, explain a systems' functioning or predict a systems' behavior and outcomes) stakeholders' actual degree and kind of understanding. This is especially important in the evaluation scenario when deficits in the substantial desiderata satisfaction become apparent. This results from the circumstance that assessing whether the required degree and kind of understanding that has been achieved allows for drawing inferences as to whether there is a fundamental gap between the explanatory process and the substantial desiderata satisfaction, or whether the provided information is not appropriate to evoke under-

standing. This is due to the fact that the explanation process can only serve to maximize understanding. If a desideratum still remains (substantially) unsatisfied, its full satisfaction goes beyond the scope of any explainability approach.

Philosophers can help to investigate and explicate what it means to have a certain degree or kind of understanding. Building on this, psychologists can design ways to empirically assess and measure such understanding (e.g., is it necessary for a given desideratum that stakeholders are able to predict a system's outputs? Is it necessary that stakeholders are able to anticipate situations when systems will likely fail?). Furthermore, scholars from law can contribute conditions for traceability and auditability.

*Explanatory Information.* The next step in the discovery scenario is to pin down what explanatory information has the potential to facilitate the right kind and degree of understanding in a predetermined context. This implies an evaluation of different dimensions of explanatory information with respect to the expected effects within an explanation process. To do so, we have to pay attention to, for instance, the kind of information, its presentation format [36], its quality [17], its amount [171], its completeness, its complexity, or its adequacy for the given context [171]. In the evaluation scenario, we have to check whether an explainability approach provides explanatory information that sufficiently meets the previously identified requirements. Sometimes there will also be a need to re-evaluate whether the requirements concerning the information are indeed adequate.

In this respect, philosophers and other explanation scientists can help to distinguish between different kinds and features of explanatory information [6,7]. Furthermore, scholars from law can examine current legislation to find out whether it prescribes certain kinds of explanatory information. In the case of the GDPR, for instance, they have to specify what it means to 'provide [...] meaningful information about the logic involved' (GDPR Art. 13 (2)(f); [101]). Finally (and based on this differentiation), educational or cognitive psychologists have the task to characterize the explanatory information that is best suited to facilitate the required kind and degree of understanding for a certain context.

*Explainability Approach.* Insights from the assessment of the concepts and relations in our model can guide and inform the requirements for explainability approaches that aim to satisfy given desiderata. This is of particular importance for the discovery scenario, where the primary objective is to identify which approach is expected to be most appropriate for providing specific information. Assume that, in the evaluation scenario we are at a point where the desiderata are not satisfied, the adequate degree and kind of understanding is not evoked, and the required explanatory information is not delivered. In this case, it is necessary to investigate whether the explainability approach is even capable of producing explanatory information with the right features at all. All the insights that are available at this point can indicate whether an existing explainability approach provides explanatory information that is sufficient to satisfy stakeholders' desiderata.

Additionally, we can learn whether a given explainability approach has the potential to derive explanatory information that can satisfy stakeholders' desiderata to a certain degree, whether it is necessary to adjust the explainability approach, whether it is sufficient to choose another one, or whether we need to develop an entirely new one. At this stage, computer scientists who can improve, adjust, and design explainability approaches are the main contributors integrating the aforementioned insights. First, they have the abilities to assess what is technically feasible and possible. Second, they can actually implement the demands regarding the explainability approach resulting from the previous steps.

## 6.2. Specific application scenario

We will conclude our thoughts about the application of our model by means of a specific example. Consider a situation where users want a system that produces fair outputs. For this, we first need an explanation process that enables them to assess whether the system produces fair outputs and, consequently, we need to find an adequate explainability approach. We first clarify the relevant user sub-classes and their prototypical characteristics. Will it be, for instance, novice or expert users? Then, we determine whether other stakeholder classes also have to be considered. For example, do we need to consider the perspectives of regulators regarding fairness? Furthermore, we have to anticipate the context. Are we talking about a personnel selection tasks or court cases with completely different contextual peculiarities?

Subsequently, we determine what users mean when they desire fair outputs (i.e., we clarify the satisfaction conditions of the desideratum's substantial facet). What kind of algorithmic fairness do they expect? At the same time, it is important to be aware of other relevant desiderata as the satisfaction of other user desiderata could be affected when trying to satisfy the fairness desideratum. Similarly, there may be unanticipated effects on the stakeholders' desiderata. For instance, will using a specific explainability approach that proves suitable to assess system fairness also affect predictive accuracy of the artificial system?

Next, we consider the desideratum's epistemic facet. Under which circumstances would users be enabled to assess whether the outputs of a system are fair? What do they need to understand with regard to a system's functioning or outputs? To answer these questions, we must determine what degree and what kind of understanding is appropriate for the epistemic satisfaction of the fairness desideratum and we need a detailed investigation of how contextual influences moderate this relation. Furthermore, we must be aware of given stakeholder characteristics (e.g., users' background knowledge).

When we have estimated what degree and kind of understanding is required to enable users to assess whether a system's outputs are fair, we can determine what kind of explanatory information facilitates this understanding. For example, we might need explanatory information about the influence of features based on protected attributes (e.g., race) on the system's outputs. Alternatively, we might need information that contrasts the treatment of different groups of people (e.g., minorities and majorities). With this knowledge, we can determine what explainability approach provides this kind of explanatory

information. Most likely, it would be post-hoc, local approaches that provide explanatory information that either highlights feature relevance or that allow users to compare subsets of instances regarding their predicted outcomes. As a result, we have specific demands regarding the explainability approach for satisfying the fairness desideratum in the case under consideration.

With this knowledge we can either choose an adequate explainability approach from existing ones or design a new one. Afterwards, we can investigate whether the explanatory information resulting from the respective explainability approach leads to a better understanding of the system and its outputs. This means that we are now in a position to empirically evaluate the success of the selected approach and the corresponding explanation process.

For this, we may conduct a stakeholder study. Given our definitions of desiderata satisfaction, we may find in this study that the epistemic facet may be satisfied (e.g., users can actually assess whether the system produces fair outputs; users can explain whether and why the respective system's outputs are fair; users can predict what kind of inputs will lead to fair or unfair outcomes) but not the substantial one (i.e., the system's outputs are actually not fair). Then, we may have to conclude that the substantial facet of the desideratum is not satisfiable by altering the explainability approach. Nevertheless, satisfying the epistemic facet of the desideratum can help us to figure out how to satisfy the substantial facet of the desideratum beyond XAI-related strategies. For instance, we may obtain information that a developer can use to improve system fairness. In this case, the artificial system has to be adjusted or changed to additionally satisfy the substantial facet of desiderata satisfaction.

On the other hand, if the stakeholder study reveals that the epistemic facet of the desideratum is not satisfied (e.g., users fail in explaining whether and why the respective system's outputs are fair), we can conclude that the users' degree or kind of understanding of the system and its outputs does not suffice. In many situations this may be the case because the explanatory information was not suitable to evoke the necessary understanding (e.g., the explainability approach did not provide explanatory information suitable to allow assessing whether the system produces fair outcomes). Depending on the context, the task, and stakeholders' individual characteristics, it may be necessary to iteratively adjust the characteristics of the explanatory information so that stakeholders engaging with this information achieve the right kind and degree of understanding. Alternatively, it may be necessary to adjust the explainability approach if it is not suitable to provide explanatory information useful for facilitating stakeholder understanding.

## 7. Conclusion

With increasing numbers of people affected by artificial systems, the number of stakeholders' desiderata will continue to grow. Although the focus of XAI research has shifted towards human stakeholders and the evaluation of explainability approaches, this shift still needs to incorporate a comprehensive view of all stakeholders and their desiderata when artificial systems are used in socially relevant contexts as well as empirical investigation of explainability approaches with respect to desiderata satisfaction. This engenders an even more pressing need for interdisciplinary collaboration in order to consider all stakeholders' perspectives and to empirically evaluate and optimally design explainability approaches to satisfy stakeholders' desiderata. The current paper has introduced a model highlighting the central concepts and their relations along which explainability approaches aim to satisfy stakeholders' desiderata. We hope that this model inspires and guides future interdisciplinary evaluation and development of explainability approaches and, thereby, further advances XAI research concerning the satisfaction of stakeholders' desiderata.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- [1] D.C. Brock, Learning from artificial intelligence's previous awakenings: the history of expert systems, *AI Mag.* 39 (3) (2018) 3–15, <https://doi.org/10.1609/aimag.v39i3.2809>.
- [2] W.J. Clancey, The epistemology of a rule-based expert system – a framework for explanation, *Artif. Intell.* 20 (3) (1983) 215–251, [https://doi.org/10.1016/0004-3702\(83\)90008-5](https://doi.org/10.1016/0004-3702(83)90008-5).
- [3] W.R. Swartout, Xplain: a system for creating and explaining expert consulting programs, *Artif. Intell.* 21 (3) (1983) 285–325, [https://doi.org/10.1016/S0004-3702\(83\)80014-9](https://doi.org/10.1016/S0004-3702(83)80014-9).
- [4] H. Johnson, P. Johnson, *Explanation facilities and interactive systems*, in: *Proceedings of the 1st International Conference on Intelligent User Interfaces*, IUI, Association for Computing Machinery, New York, NY, USA, 1993, pp. 159–166.
- [5] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*, XAI, 2017, pp. 8–13.
- [6] T. Miller, Explanation in artificial intelligence: insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38, <https://doi.org/10.1016/j.artint.2018.07.007>.
- [7] B.D. Mittelstadt, C. Russell, S. Wachter, *Explaining explanations in AI*, in: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 279–288.
- [8] J. Burrell, How the machine 'thinks': understanding opacity in machine learning algorithms, *Big Data Soc.* 3 (1) (2016) 1–12, <https://doi.org/10.1177/2053951715622512>.
- [9] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, *CoRR*, arXiv:1702.08608 [abs], 2017.

- [10] EU High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, 2019.
- [11] Z.C. Lipton, The mythos of model interpretability, *Commun. ACM* 61 (10) (2018) 36–43, <https://doi.org/10.1145/3233231>.
- [12] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160, <https://doi.org/10.1109/ACCESS.2018.2870052>.
- [13] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, *User Model. User-Adapt. Interact.* 27 (3–5) (2017) 393–444, <https://doi.org/10.1007/s11257-017-9195-0>.
- [14] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115, <https://doi.org/10.1016/j.inffus.2019.12.012>.
- [15] H. Felzmann, E.F. Villaronga, C. Lutz, A. Tamò-Larrieux, Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns, *Big Data Soc.* 6 (1) (2019) 1–14, <https://doi.org/10.1177/2053951719860542>.
- [16] L.H. Gilpin, C. Testart, N. Fruchter, J. Adebayo, Explaining explanations to society, in: *NIPS Workshop on Ethical, Social and Governance Issues in AI*, 2018, pp. 1–6, arXiv:1901.06560 [abs].
- [17] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, *CoRR*, arXiv:1812.04608 [abs], 2018.
- [18] A.D. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable AI, *CoRR*, arXiv:1810.00184 [abs], 2018.
- [19] A. Weller, Transparency: motivations and challenges, in: W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019, pp. 23–40.
- [20] A. Páez, The pragmatic turn in Explainable Artificial Intelligence (XAI), *Minds Mach.* 29 (2019) 441–459, <https://doi.org/10.1007/s11023-019-09502-w>.
- [21] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F.M. Harper, H. Zhu, Explaining decision-making algorithms through ui: strategies to help non-expert stakeholders, in: *Proceedings of the 2019 chi Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–12.
- [22] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda, in: *Proceedings of the 2018 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–18.
- [23] M.-W. Dictionary, Stakeholder, <https://www.merriam-webster.com/dictionary/stakeholder>, 2020. (Accessed 30 July 2020).
- [24] M. Hind, D. Wei, M. Campbell, N.C.F. Codella, A. Dhurandhar, A. Mojsilović, K.N. Ramamurthy, K.R. Varshney, Ted: teaching AI to explain its decisions, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 123–129.
- [25] S. Anjomshoei, K. Främling, A. Najjar, Explanations of black-box model predictions by contextual importance and utility, in: *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer, 2019, pp. 95–109.
- [26] M. Atzmüller, Towards socio-technical design of explicative systems: transparent, interpretable and explainable analytics and its perspectives in social interaction contexts information, in: *Proceedings of the 2019 Workshop on Affective Computing and Context Awareness in Ambient Intelligence*, AfCAI, 2019, pp. 1–8.
- [27] I. Baaj, J.-P. Poli, W. Ouerdane, Some insights towards a unified semantic representation of explanation for explainable artificial intelligence, in: *Proceedings of the 2019 Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, NL4XAI, Association for Computational Linguistics, 2019, pp. 14–19.
- [28] K. Balog, F. Radlinski, S. Arakelyan, Transparent, scrutable and explainable user models for personalized recommendation, in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 265–274.
- [29] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions, in: *Proceedings of the 2018 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–14.
- [30] T. Chakraborti, S. Sreedharan, S. Grover, S. Kambhampati, Plan explanations as model reconciliation, in: *14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI, IEEE, 2019, pp. 258–266.
- [31] L. Chen, D. Yan, F. Wang, User evaluations on sentiment-based recommendation explanations, *ACM Trans. Interact. Intell. Syst.* 9 (4) (2019) 1–38, <https://doi.org/10.1145/3282878>.
- [32] K. Cotter, J. Cho, E. Rader, Explaining the news feed algorithm: an analysis of the "news feed fy" blog, in: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1553–1560.
- [33] K. Darlington, Aspects of intelligent systems explanation, *Univers. J. Control Autom.* 1 (2) (2013) 40–51, <https://doi.org/10.13189/ujca.2013.010204>.
- [34] K. Ehrlich, S.E. Kirk, J. Patterson, J.C. Rasmussen, S.I. Ross, D.M. Gruen, Taking advice from intelligent systems: the double-edged sword of explanations, in: *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI, Association for Computing Machinery, New York, NY, USA, 2011, pp. 125–134.
- [35] A.A. Freitas, Comprehensible classification models: a position paper, *ACM SIGKDD Explor. Newsl.* 15 (1) (2014) 1–10, <https://doi.org/10.1145/2594473.2594475>.
- [36] S. Gregor, I. Benbasat, Explanations from intelligent systems: theoretical foundations and implications for practice, *MIS Q.* 23 (4) (1999) 497–530, <https://doi.org/10.2307/249487>.
- [37] J. Hois, D. Theofanou-Fuelbier, A.J. Junk, How to achieve explainability and transparency in human AI interaction, in: *International Conference on Human-Computer Interaction*, HCI, Springer, 2019, pp. 177–183.
- [38] R.F. Kizilcec, How much information? effects of transparency on trust in an algorithmic interface, in: *Proceedings of the 2016 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2016, pp. 2390–2395.
- [39] S. Nagulendra, J. Vassileva, Providing awareness, explanation and control of personalized filtering in a social networking site, *Inf. Syst. Front.* 18 (1) (2016) 145–158, <https://doi.org/10.1007/s10796-015-9577-y>.
- [40] A. Papenmeier, G. Englebienne, C. Seifert, How model accuracy and explanation fidelity influence user trust in AI, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 94–100.
- [41] R. Pierrard, J.-P. Poli, C. Hudelot, A new approach for explainable multiple organ annotation with few data, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 101–107.
- [42] V. Putnam, L. Riegel, C. Conati, Towards personalized XAI: a case study in intelligent tutoring systems, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 108–114.
- [43] E. Rader, K. Cotter, J. Cho, Explanations as mechanisms for supporting algorithmic transparency, in: *Proceedings of the 2018 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–13.
- [44] A. Rosenfeld, A. Richardson, Explainability in human–agent systems, *Auton. Agents Multi-Agent Syst.* 33 (6) (2019) 673–705, <https://doi.org/10.1007/s10458-019-09408-y>.
- [45] M. Sato, K. Nagatani, T. Sonoda, Q. Zhang, T. Ohkuma, Context style explanation for recommender systems, *J. Inf. Process.* 27 (2019) 720–729, <https://doi.org/10.2197/ipsjjip.27.720>.

- [46] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI, Association for Computing Machinery, New York, NY, USA, 2009, pp. 47–56.
- [47] X. Watts, F. Lécué, Local score dependent model explanation for time dependent covariates, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 129–135.
- [48] J. Zhou, H. Hu, Z. Li, K. Yu, F. Chen, Physiological indicators for user trust in machine learning with influence enhanced fact-checking, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2019, pp. 94–113.
- [49] J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW, Association for Computing Machinery, New York, NY, USA, 2000, pp. 241–250.
- [50] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The effects of transparency on trust in and acceptance of a content-based art recommender, *User Model. User-Adapt. Interact.* 18 (5) (2008) 455, <https://doi.org/10.1007/s11257-008-9051-3>.
- [51] R.M. Byrne, Counterfactuals in Explainable Artificial Intelligence (XAI): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282.
- [52] P.B. De Laat, Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?, *Philos. Technol.* 31 (4) (2018) 525–541, <https://doi.org/10.1007/s13347-017-0293-z>.
- [53] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations, *Minds Mach.* 28 (4) (2018) 689–707, <https://doi.org/10.1007/s11023-018-9482-5>.
- [54] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, P. Vinck, Fair, transparent, and accountable algorithmic decision-making processes, *Philos. Technol.* 31 (4) (2018) 611–627, <https://doi.org/10.1007/s13347-017-0279-x>.
- [55] S.M. Mathews, Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review, in: *Intelligent Computing – Proceedings of the Computing Conference*, Springer, 2019, pp. 1269–1292.
- [56] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: mapping the debate, *Big Data Soc.* 3 (2) (2016) 1–21, <https://doi.org/10.1177/2053951716679679>.
- [57] W. Pieters, Explanation and trust: what to tell the user in security and AI?, *Ethics Inf. Technol.* 13 (1) (2011) 53–64, <https://doi.org/10.1007/s10676-010-9253-3>.
- [58] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: users, values, concerns and challenges, in: *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 2018, pp. 19–36.
- [59] M.O. Riedl, Human-centered artificial intelligence and machine learning, *Hum. Behav. Emerg. Technol.* 1 (1) (2019) 33–36, <https://doi.org/10.1002/hbe2.117>.
- [60] S. Robbins, A misdirected principle with a catch: explicability for AI, *Minds Mach.* 29 (4) (2019) 495–514, <https://doi.org/10.1007/s11023-019-09509-3>.
- [61] R. Sheh, Different XAI for different HRI, in: *AAAI Fall Symposium*, AAAI Press, 2017, pp. 114–117.
- [62] R. Sheh, I. Monteath, Defining explainable AI for requirements analysis, *Künstl. Intell.* 32 (4) (2018) 261–266, <https://doi.org/10.1007/s13218-018-0559-3>.
- [63] K. Sokol, P.A. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 56–67.
- [64] K. Sokol, P. Flach, One explanation does not fit all, *Künstl. Intell.* 34 (2) (2020) 235–250, <https://doi.org/10.1007/s13218-020-00637-y>.
- [65] M. Sridharan, B. Meadows, Towards a theory of explanations for human–robot collaboration, *Künstl. Intell.* 33 (4) (2019) 331–342, <https://doi.org/10.1007/s13218-019-00616-y>.
- [66] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, *Neural Comput. Appl.* (2019), <https://doi.org/10.1007/s00521-019-04051-w>.
- [67] D. Wang, Q. Yang, A. Abdul, B.Y. Lim, Designing theory-driven user-centric explainable AI, in: *Proceedings of the 2019 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15.
- [68] M.K. Lee, A. Jain, H.J. Cha, S. Ojha, D. Kusbit, Procedural justice in algorithmic fairness, *Proc. ACM Human-Comput. Interact.* 3 (2019) 1–26, <https://doi.org/10.1145/3359284>.
- [69] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? a new conceptualization of perspectives, in: *CEUR Workshop Proceedings*, vol. 2071, CEUR, 2018, pp. 1–8, arXiv:1710.00794.
- [70] M. Krishnan, Against interpretability: a critical examination of the interpretability problem in machine learning, *Philos. Technol.* (2019) 1–16, <https://doi.org/10.1007/s13347-019-00372-9>.
- [71] S.K. Peddalu, M. Saravanan, S. Suresh, Explainable classification using clustering in deep learning models, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 115–121.
- [72] N.F. Rajani, R.J. Mooney, Using explanations to improve ensembling of visual question answering systems, in: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*, XAI, 2017, pp. 43–47.
- [73] J. Zhou, F. Chen, Towards trustworthy human–AI teaming under uncertainty, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 143–147.
- [74] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: results from a systematic literature review, in: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, 2019, pp. 1078–1088.
- [75] M. Fox, D. Long, D. Magazzeni, Explainable planning, in: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*, XAI, 2017, pp. 24–30.
- [76] S. Jasanoft, Virtual, visible, and actionable: data assemblages and the sightlines of justice, *Big Data Soc.* 4 (2) (2017) 1–15, <https://doi.org/10.1177/2053951717724477>.
- [77] G. Friedrich, M. Zanker, A taxonomy for generating explanations in recommender systems, *AI Mag.* 32 (3) (2011) 90–98, <https://doi.org/10.1609/aimag.v32i3.2365>.
- [78] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 9 (4) (2019) 1–13, <https://doi.org/10.1002/widm.1312>.
- [79] R. Sevastjanova, F. Beck, B. Ell, C. Turky, R. Henkin, M. Butt, D.A. Keim, M. El-Assady, Going beyond visualization: verbalization as complementary medium to explain machine learning models, in: *VIS Workshop on Visualization for AI Explainability*, VISxAI, 2018, pp. 1–6.
- [80] F. Sørmo, J. Cassens, A. Aamodt, Explanation in case-based reasoning—perspectives and goals, *Artif. Intell. Rev.* 24 (2) (2005) 109–143, <https://doi.org/10.1007/s10462-005-4607-7>.
- [81] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Transparency in algorithmic and human decision-making: is there a double standard?, *Philos. Technol.* 32 (4) (2019) 661–683, <https://doi.org/10.1007/s13347-018-0330-6>.
- [82] A. Lucic, H. Haned, M. de Rijke, Contrastive explanations for large errors in retail forecasting predictions through Monte Carlo simulations, in: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, XAI, 2019, pp. 66–72.
- [83] H.K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, ICSE-NIER, Association for Computing Machinery, New York, NY, USA, 2018, pp. 53–56.

- [84] J. De Winter, Explanations in software engineering: the pragmatic point of view, *Minds Mach.* 20 (2) (2010) 277–289, <https://doi.org/10.1007/s11023-010-9190-2>.
- [85] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, F. Doshi-Velez, Explainable reinforcement learning via reward decomposition, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 47–53.
- [86] L. Michael, Machine coaching, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 80–86.
- [87] K. Sokol, P.A. Flach, Conversational explanations of machine learning predictions through class-contrastive counterfactual statements, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 2018, pp. 5785–5786.
- [88] H. Wicaksono, C. Sammut, R. Sheh, Towards explainable tool creation by a robot, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 63–67.
- [89] T. Eiter, Z.G. Saribatur, P. Schüller, Abstraction for zooming-in to unsolvability reasons of grid-cell problems, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 7–13.
- [90] T. Kulesza, S. Stumpf, W.-K. Wong, M.M. Burnett, S. Perona, A. Ko, I. Oberst, Why-oriented end-user debugging of naive Bayes text classification, *ACM Trans. Interact. Intell. Syst.* 1 (1) (2011) 1–31, <https://doi.org/10.1145/2030365.2030367>.
- [91] R.R. Hoffman, G. Klein, S.T. Mueller, Explaining explanation for “explainable AI”, *Proc. Hum. Factors Ergon Soc. Ann. Meet.* 62 (2018) 197–201, <https://doi.org/10.1177/1541931218621047>.
- [92] F. Nothdurft, T. Heinroth, W. Minker, The impact of explanation dialogues on human-computer trust, in: International Conference on Human-Computer Interaction, HCI, Springer, 2013, pp. 59–67.
- [93] C. Brinton, A framework for explanation of machine learning decisions, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 14–18.
- [94] N. Tintarev, Explanations of recommendations, in: Proceedings of the 2007 ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, 2007, pp. 203–206.
- [95] R.O. Weber, H. Hong, P. Goel, Explaining citation recommendations: abstracts or full texts?, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 136–142.
- [96] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning, in: IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA, 2018, pp. 80–89.
- [97] J.C.-T. Ho, How biased is the sample? reverse engineering the ranking algorithm of facebook’s graph application programming interface, *Big Data Soc.* 7 (1) (2020) 1–15, <https://doi.org/10.1177/2053951720905874>.
- [98] F. Hohman, A. Head, R. Caruana, R. DeLine, S.M. Drucker, Gamut: a design probe to understand how data scientists understand machine learning models, in: Proceedings of the 2019 Conference on Human Factors in Computing Systems, CHI, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13.
- [99] M. Veale, R. Binns, Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data, *Big Data Soc.* 4 (2) (2017) 1–17, <https://doi.org/10.1177/2053951717743530>.
- [100] C. Zednik, Solving the black box problem: a normative framework for explainable artificial intelligence, *Philos. Technol.* (2019) 1–24, <https://doi.org/10.1007/s13347-019-00382-7>.
- [101] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a “right to explanation”, *AI Mag.* 38 (3) (2017) 50–57, <https://doi.org/10.1609/aimag.v38i3.2741>.
- [102] E.I. Sklar, M.Q. Azhar, Explanation through argumentation, in: Proceedings of the 6th International Conference on Human-Agent Interaction, HAI, Association for Computing Machinery, New York, NY, USA, 2018, pp. 277–285.
- [103] I. Lage, D. Lifschitz, F. Doshi-Velez, O. Amir, Exploring computational user models for agent policy summarization, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 59–65.
- [104] E.S. Dahl, Appraising black-boxed technology: the positive prospects, *Philos. Technol.* 31 (4) (2018) 571–591, <https://doi.org/10.1007/s13347-017-0275-1>.
- [105] B. Ghosh, D. Malioutov, K.S. Meel, Interpretable classification rules in relaxed logical form, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 14–20.
- [106] M.T. Stuart, N.J. Nersessian, Peeking inside the black box: a new kind of scientific visualization, *Minds Mach.* 29 (1) (2019) 87–107, <https://doi.org/10.1007/s11023-018-9484-3>.
- [107] J. Clos, N. Wiratunga, S. Massie, Towards explainable text classification by jointly learning lexicon and modifier terms, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 19–23.
- [108] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G.M. Youngblood, Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation, in: IEEE Conference on Computational Intelligence and Games, CIG, IEEE, 2018, pp. 1–8.
- [109] M.-A. Clinciu, H. Hastie, A survey of explainable AI terminology, in: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, NL4XAI 2019, Association for Computational Linguistics, 2019, pp. 8–13.
- [110] C. Henin, D. Le Métayer, Towards a generic framework for black-box explanation methods, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 28–34.
- [111] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, 2019, pp. 1033–1041.
- [112] M.L. Olson, L. Neal, F. Li, W.-K. Wong, Counterfactual states for Atari agents via generative deep learning, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 87–93.
- [113] Z. Zeng, C. Miao, C. Leung, J.J. Chin, Building more explainable artificial intelligence with argumentation, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18, AAAI Press, 2018, pp. 8044–8046, <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16762>.
- [114] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 73–79.
- [115] M.T. Ribeiro, S. Singh, C. Guestrin, “Why should I trust you?”: explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144.
- [116] M.R. Endsley, From here to autonomy, *Hum. Factors, J. Hum. Factors Ergon. Soc.* 59 (1) (2017) 5–27, <https://doi.org/10.1177/0018720816681350>.
- [117] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, *Hum. Factors* 46 (1) (2004) 50–80, [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392).
- [118] R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse, *Hum. Factors, J. Hum. Factors Ergon. Soc.* 39 (2) (1997) 230–253, <https://doi.org/10.1518/001872097778543886>.
- [119] K.A. Hoff, M. Bashir, Trust in automation, *Hum. Factors* 57 (3) (2014) 407–434, <https://doi.org/10.1177/0018720814547570>.
- [120] R. Parasuraman, D.H. Manzey, Complacency and bias in human use of automation: an attentional integration, *Hum. Factors* 52 (3) (2010) 381–410, <https://doi.org/10.1177/0018720810376055>.
- [121] A. Kunze, S.J. Summerskill, R. Marshall, A.J. Filtness, Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces, *Ergonomics* 62 (3) (2019) 345–360, <https://doi.org/10.1080/00140139.2018.1547842>.

- [122] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, CoRR, arXiv: 1708.08296 [abs], 2017.
- [123] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, *Digit. Signal Process.* 73 (2018) 1–15, <https://doi.org/10.1016/j.dsp.2017.10.011>.
- [124] S. Becker, M. Ackermann, S. Lapuschkin, K. Müller, W. Samek, Interpreting and explaining deep neural networks for classification of audio signals, CoRR, arXiv:1807.03418 [abs], 2018.
- [125] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek, Analyzing classifiers: fisher vectors and deep neural networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2912–2920.
- [126] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1721–1730.
- [127] K. Baum, H. Hermanns, T. Speith, From machine ethics to machine explainability and back, in: *International Symposium on Artificial Intelligence and Mathematics*, ISAIM, 2018, pp. 1–8, [http://isaim2018.cs.virginia.edu/papers/ISAIM2018\\_Ethics\\_Baum\\_et.al.pdf](http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Ethics_Baum_et.al.pdf).
- [128] C. Luetge, The german ethics code for automated and connected driving, *Philos. Technol.* 30 (2017) 547–558, <https://doi.org/10.1007/s13347-017-0284-0>.
- [129] S.I.S. Purkiss, P.L. Perrewé, T.L. Gillespie, B.T. Mayes, G.R. Ferris, Implicit sources of bias in employment interview judgments and decisions, *Organ. Behav. Hum. Decis. Process.* 101 (2) (2006) 152–167, <https://doi.org/10.1016/j.obhdp.2006.06.005>.
- [130] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* 356 (2017) 183–186, <https://doi.org/10.1126/science.aal4230>.
- [131] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (5) (2019) 1–42, <https://doi.org/10.1145/3236009>.
- [132] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: toward a unified view, *Manag. Inf. Syst. Q.* 27 (3) (2003) 425–478, <https://doi.org/10.2307/30036540>.
- [133] C. McLeod, Trust, in: E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, fall 2015 edition, Metaphysics Research Lab., Stanford University, 2015, pp. 1–43.
- [134] I.D. Raji, A. Smart, R.N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT\*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 33–44.
- [135] A. Matthias, The responsibility gap: ascribing responsibility for the actions of learning automata, *Ethics Inf. Technol.* 6 (3) (2004) 175–183, <https://doi.org/10.1007/s10676-004-3422-1>.
- [136] E.L. Deci, A.H. Olafsen, R.M. Ryan, Self-determination theory in work organizations: the state of a science, *Ann. Rev. Organ. Psychol. Organ. Behav.* 4 (1) (2017) 19–43, <https://doi.org/10.1146/annurev-orgpsych-032516-113108>.
- [137] C. Longoni, A. Bonezzi, C.K. Morewedge, Resistance to medical artificial intelligence, *J. Consum. Res.* 46 (4) (2019) 629–650, <https://doi.org/10.1093/jcr/ucz013>.
- [138] F.C. Keil, Explanation and understanding, *Annu. Rev. Psychol.* 57 (1) (2006) 227–254, <https://doi.org/10.1146/annurev.psych.57.102904.190100>.
- [139] J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles, *Science* 352 (2016) 1573–1576, <https://doi.org/10.1126/science.aaf2654>.
- [140] B.G. Buchanan, E.H. Shortliffe, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, 1984.
- [141] J.S. Dhaliwal, I. Benbasat, The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation, *Inf. Syst. Res.* 7 (3) (1996) 342–362, <https://doi.org/10.1287/isre.7.3.342>.
- [142] M.A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, D. Bohlander, Explainability as a non-functional requirement, in: *IEEE 27th International Requirements Engineering Conference*, RE, 2019, pp. 363–368.
- [143] H.W. De Regt, *Understanding Scientific Understanding*, Oxford University Press, 2017.
- [144] C. Baumberger, C. Beisbart, G. Brun, What is understanding? An overview of recent debates in epistemology and philosophy of science, in: S.G.C. Baumberger, S. Ammon (Eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Routledge, 2017, pp. 1–34.
- [145] F.I. Malfatti, On understanding and testimony, *Erkenntnis* (2019) 1–21, <https://doi.org/10.1007/s10670-019-00157-8>.
- [146] C. Baumberger, Types of understanding: their nature and their relation to knowledge, *Conceptus* 40 (98) (2014) 67–88, <https://doi.org/10.1515/cpt-2014-0002>.
- [147] K. Lambert, On whether an answer to a why-question is an explanation if and only if it yields scientific understanding, in: G.G. Brittan (Ed.), *Causality, Method, and Modality*, vol. 48, Springer, Dordrecht, Netherlands, 1991, pp. 125–142.
- [148] T. Lombrozo, S. Carey, Functional explanation and the function of explanation, *Cognition* 99 (2) (2006) 167–204, <https://doi.org/10.1016/j.cognition.2004.12.009>.
- [149] M.T. Chi, N. De Leeuw, M.-H. Chiu, C. Lavancher, Eliciting self-explanations improves understanding, *Cogn. Sci.* 18 (3) (1994) 439–477, [https://doi.org/10.1207/s15516709cog1803\\_3](https://doi.org/10.1207/s15516709cog1803_3).
- [150] R.E. Mayer, Cognition and instruction: their historic meeting within educational psychology, *J. Educ. Psychol.* 84 (4) (1992) 405–412, <https://doi.org/10.1037/0022-0663.84.4.405>.
- [151] S.T. Mueller, R.R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, CoRR, arXiv:1902.01876 [abs], 2019.
- [152] C. Kelp, Understanding phenomena, *Synthese* 192 (12) (2015) 3799–3816, <https://doi.org/10.1007/s11229-014-0616-x>.
- [153] P.J. Feltovich, R.L. Coulson, R.J. Spiro, Learners' (mis)understanding of important and difficult concepts: a challenge to smart machines in education, in: *Smart Machines in Education: The Coming Revolution in Educational Technology*, MIT Press, Cambridge, MA, USA, 2001, pp. 349–375.
- [154] W.B. Rouse, N.M. Morris, On looking into the black box: prospects and limits in the search for mental models, *Psychol. Bull.* 100 (3) (1986) 349–363, <https://doi.org/10.1037/0033-2959.100.3.349>.
- [155] L. Rozenblit, F. Keil, The misunderstood limits of folk science: an illusion of explanatory depth, *Cogn. Sci.* 26 (5) (2002) 521–562, [https://doi.org/10.1207/s15516709cog2605\\_1](https://doi.org/10.1207/s15516709cog2605_1).
- [156] D. Kuhn, How do people know?, *Psychol. Sci.* 12 (1) (2001) 1–8, <https://doi.org/10.1111/1467-9280.00302>.
- [157] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: *IEEE Symposium on Visual Languages and Human-Centric Computing*, IEEE, 2013, pp. 3–10.
- [158] J. Tullio, A.K. Dey, J. Chalecki, J. Fogarty, How it works: a field study of non-technical users interacting with an intelligent system, in: *Proceedings of the 2007 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2007, pp. 31–40.
- [159] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model cards for model reporting, in: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 220–229.
- [160] M. Langer, C.J. König, A. Fitili, Information as a double-edged sword: the role of computer experience and information on applicant reactions towards novel technologies for personnel selection, *Comput. Hum. Behav.* 81 (2018) 19–30, <https://doi.org/10.1016/j.chb.2017.11.036>.

- [161] D.T. Newman, N.J. Fast, D.J. Harmon, When eliminating bias isn't fair: algorithmic reductionism and procedural justice in human resource decisions, *Organ. Behav. Hum. Decis. Process.* 160 (2020) 149–167, <https://doi.org/10.1016/j.obhdp.2020.03.008>.
- [162] M. Bazire, P. Brézillon, *Understanding context before using it*, in: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), *Modeling and Using Context*, Springer, 2005, pp. 29–40.
- [163] P. Dourish, What we talk about when we talk about context, *Pers. Ubiquitous Comput.* 8 (1) (2004) 19–30, <https://doi.org/10.1007/s00779-003-0253-8>.
- [164] D.R. Bobocel, A. Zdaniuk, How can explanations be used to foster organizational justice, in: *Handbook of Organizational Justice*, 2005, pp. 469–498.
- [165] R. Folger, R. Cropanzano, Fairness theory: justice as accountability, in: *Advances in Organizational Justice*, vol. 1, 2001, pp. 1–55.
- [166] J.C. Shaw, E. Wild, J.A. Colquitt, To justify or excuse?: A meta-analytic review of the effects of explanations, *J. Appl. Psychol.* 88 (3) (2003) 444–458, <https://doi.org/10.1037/0021-9010.88.3.444>.
- [167] J. Brockner, B.M. Wiesenfeld, An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures, *Psychol. Bull.* 120 (2) (1996) 189–208, <https://doi.org/10.1037/0033-2909.120.2.189>.
- [168] R. Wang, F.M. Harper, H. Zhu, Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences, in: *Proceedings of the 2020 Conference on Human Factors in Computing Systems*, CHI, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14.
- [169] E.A. Lind, K. van den Bos, When fairness works: toward a general theory of uncertainty management, *Res. Organ. Behav.* 24 (2002) 181–223, [https://doi.org/10.1016/s0191-3085\(02\)24006-x](https://doi.org/10.1016/s0191-3085(02)24006-x).
- [170] J.A. Colquitt, J.M. Chertkoff, Explaining injustice: the interactive effect of explanation and outcome on fairness perceptions and task motivation, *J. Manag.* 28 (5) (2002) 591–610, <https://doi.org/10.1177/014920630202800502>.
- [171] P. Liu, Z. Li, Task complexity: a review and conceptualization framework, *Int. J. Ind. Ergon.* 42 (6) (2012) 553–568, <https://doi.org/10.1016/j.ergon.2012.09.001>.
- [172] D.A. Wilkenfeld, Functional explaining: a new approach to the philosophy of explanation, *Synthese* 191 (2014) 3367–3391, <https://doi.org/10.1007/s11229-014-0452-z>.
- [173] D.A. Wilkenfeld, D. Plunkett, T. Lombrozo, Depth and deference: when and why we attribute understanding, *Philos. Stud.* 173 (2016) 373–393, <https://doi.org/10.1007/s11098-015-0497-y>.
- [174] T. Lombrozo, The instrumental value of explanations, *Philos. Compass* 6 (8) (2011) 539–551, <https://doi.org/10.1111/j.1747-9991.2011.00413.x>.
- [175] J.J. Williams, T. Lombrozo, Explanation and prior knowledge interact to guide learning, *Cogn. Psychol.* 66 (1) (2013) 55–84, <https://doi.org/10.1016/j.cogpsych.2012.09.002>.
- [176] T. Lombrozo, B. Rehder, Functions in biological kind classification, *Cogn. Psychol.* 65 (4) (2012) 457–485, <https://doi.org/10.1016/j.cogpsych.2012.06.002>.
- [177] C.G. Hempel, Deductive-nomological explanation, in: *Aspects of Scientific Explanation*, Free Press, 1965, pp. 335–376.
- [178] W.C. Salmon, *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, 1984.
- [179] P. Gärdenfors, *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, MIT Press, 1988.
- [180] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, *Harv. J. Law Technol.* 31 (2) (2018), <https://doi.org/10.2139/ssrn.3063289>.
- [181] C.F. Craver, *Explaining the Brain*, Oxford University Press, Oxford, 2007.
- [182] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd edition, Cambridge University Press, 2009.
- [183] P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction and Search*, 2nd edition, MIT Press, 2001.
- [184] D. Borsboom, A. Cramer, A. Kalis, Brain disorders? Not really ... why network structures block reductionism in psychopathology research, *Behav. Brain Sci.* 42 (2018) 1–54, <https://doi.org/10.1017/S0140525X17002266>.
- [185] T. Lombrozo, Simplicity and probability in causal explanation, *Cogn. Psychol.* 55 (3) (2007) 232–257, <https://doi.org/10.1016/j.cogpsych.2006.09.006>.
- [186] N. Vasilyeva, D. Wilkenfeld, T. Lombrozo, Contextual utility affects the perceived quality of explanations, *Psychon. Bull. Rev.* 24 (2017) 1436–1450, <https://doi.org/10.3758/s13423-017-1275-y>.
- [187] V. Bellotti, K. Edwards, Intelligibility and accountability: human considerations in context-aware systems, *Hum.-Comput. Interact.* 16 (2–4) (2001) 193–212, [https://doi.org/10.1207/s15327051hci16234\\_05](https://doi.org/10.1207/s15327051hci16234_05).
- [188] K. Hartley, L.D. Bendixen, Educational research in the internet age: examining the role of individual characteristics, *Educ. Res.* 30 (9) (2001) 22–26, <https://doi.org/10.3102/0013189X030009022>.
- [189] H. Kauffman, A review of predictive factors of student success in and satisfaction with online learning, *Res. Learn. Technol.* 23 (2015), <https://doi.org/10.3402/rlt.v23.26507>.
- [190] D.S. McNamara, E. Kintsch, N.B. Songer, W. Kintsch, Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text, *Cogn. Instr.* 14 (1) (1996) 1–43, [https://doi.org/10.1207/s1532690xci1401\\_1](https://doi.org/10.1207/s1532690xci1401_1).
- [191] L. Goldberg, *Language and individual differences: the search for universals in personality lexicons*, in: L. Wheeler (Ed.), *Review of Personality and Social Psychology*, vol. 2, Edition, SAGE Publications, 1981, pp. 141–166.
- [192] J.T. Cacioppo, R.E. Petty, The need for cognition, *J. Pers. Soc. Psychol.* 42 (1) (1982) 116–131, <https://doi.org/10.1037/0022-3514.42.1.116>.
- [193] C.P. Haugvedt, R.E. Petty, Personality and persuasion: need for cognition moderates the persistence and resistance of attitude changes, *J. Pers. Soc. Psychol.* 63 (2) (1992) 308–319, <https://doi.org/10.1037/0022-3514.63.2.308>.
- [194] T.K. DeBacker, H.M. Crowson, The influence of need for closure on learning and teaching, *Educ. Psychol. Rev.* 21 (2009) 303–323, <https://doi.org/10.1007/s10648-009-9111-1>.
- [195] D.M. Webster, A.W. Kruglanski, Individual differences in need for cognitive closure, *J. Pers. Soc. Psychol.* 67 (6) (1994) 1049–1062, <https://doi.org/10.1037/0022-3514.67.6.1049>.
- [196] P.M. Fernbach, S.A. Sloman, R.S. Louis, J.N. Shube, Explanation fiends and foes: how mechanistic detail determines understanding and preference, *J. Consum. Res.* 39 (5) (2012) 1115–1131, <https://doi.org/10.1086/667782>.
- [197] L. Hasher, R.T. Zacks, *Working memory, comprehension, and aging: a review and a new view*, in: *Psychology of Learning and Motivation*, Elsevier, 1988, pp. 193–225.
- [198] R. Ackerman, T. Lauterman, Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure, *Comput. Hum. Behav.* 28 (5) (2012) 1816–1828, <https://doi.org/10.1016/j.chb.2012.04.023>.
- [199] M.S. Prewett, R.C. Johnson, K.N. Saboe, L.R. Elliott, M.D. Coover, Managing workload in human–robot interaction: a review of empirical studies, *Comput. Hum. Behav.* 26 (5) (2010) 840–856, <https://doi.org/10.1016/j.chb.2010.03.010>.
- [200] K. Starcke, O.T. Wolf, H.J. Markowitzsch, M. Brand, Anticipatory stress influences decision making under explicit risk conditions, *Behav. Neurosci.* 122 (6) (2008) 1352–1360, <https://doi.org/10.1037/a0013281>.
- [201] S. Lupien, F. Maheu, M. Tu, A. Fiocco, T. Schramek, The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition, *Brain Cogn.* 65 (3) (2007) 209–237, <https://doi.org/10.1016/j.bandc.2007.02.007>.
- [202] P.A. Hancock, On the process of automation transition in multitask human–machine systems, *IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum.* 37 (4) (2007) 586–598, <https://doi.org/10.1109/tsmca.2007.897610>.
- [203] L. Chazette, O. Karras, K. Schneider, Do end-users want explanations? Analyzing the role of explainability as an emerging aspect of non-functional requirements, in: *IEEE 27th International Requirements Engineering Conference*, RE, 2019, pp. 223–233.

- [204] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, *Requir. Eng.* 25 (4) (2020) 493–514, <https://doi.org/10.1007/s00766-020-00333-1>.
- [205] V. Arya, R.K.E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q.V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J.T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K.R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques, *CoRR*, arXiv:1909.03012 [abs], 2019.
- [206] J. Woodward, Scientific explanation, in: E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, winter 2019 edition, Metaphysics Research Lab., Stanford University, 2019, pp. 1–101.
- [207] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, A. Preece, A systematic method to understand requirements for explainable AI (XAI) systems, in: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, XAI, 2019, pp. 21–27.
- [208] T. Miller, P. Howe, L. Sonenberg, Explainable AI: beware of inmates running the asylum, or: how I learnt to stop worrying and love the social and behavioural sciences, in: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, 2017, pp. 36–42.
- [209] B. Kim, C. Rudin, J.A. Shah, The Bayesian case model: a generative approach for case-based reasoning and prototype classification, in: *Advances in Neural Information Processing Systems*, 2014, pp. 1952–1960.
- [210] B. Kim, R. Khanna, S. Koyejo, Examples are not enough, learn to criticize! Criticism for interpretability, in: *Advances in Neural Information Processing Systems*, 2016, pp. 2280–2288.
- [211] J.M. Carroll, M.B. Rosson, Paradox of the active user, in: *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*, MIT Press, Cambridge, MA, USA, 1987, pp. 80–111.



# From Responsibility to Reason-Giving Explainable Artificial Intelligence

Kevin Baum<sup>1,2</sup> · Susanne Mantel<sup>1</sup> · Eva Schmidt<sup>3</sup> · Timo Speith<sup>1,2</sup>

Received: 2 July 2021 / Accepted: 20 November 2021 / Published online: 19 February 2022  
© The Author(s) 2022

## Abstract

We argue that explainable artificial intelligence (XAI), specifically reason-giving XAI, often constitutes the most suitable way of ensuring that someone can properly be held responsible for decisions that are based on the outputs of artificial intelligent (AI) systems. We first show that, to close moral responsibility gaps (Matthias 2004), often a human in the loop is needed who is directly responsible for particular AI-supported decisions. Second, we appeal to the epistemic condition on moral responsibility to argue that, in order to be responsible for her decision, the human in the loop has to have an explanation available of the system's recommendation. Reason explanations are especially well-suited to this end, and we examine whether—and how—it might be possible to make such explanations fit with AI systems. We support our claims by focusing on a case of disagreement between human in the loop and AI system.

**Keywords** Explainable artificial intelligence · Reasons · Reason explanations · Moral responsibility · Responsibility gap · Decision support systems

---

This article is part of the Topical Collection on *AI and Responsibility*.

---

All authors share first authorship equally.

Eva Schmidt  
eva.schmidt@tu-dortmund.de

Kevin Baum  
kevin.baum@uni-saarland.de

Susanne Mantel  
susanne.mantel@uni-saarland.de

Timo Speith  
timo.speith@uni-saarland.de

<sup>1</sup> Institute of Philosophy, Saarland University, Saarbrücken, Germany

<sup>2</sup> Department of Computer Science, Saarland University, Saarbrücken, Germany

<sup>3</sup> Department of Philosophy and Political Science, Emil-Figge-Str. 50, 44227 TU Dortmund, Germany

## 1 Introduction

Sophisticated artificially intelligent (AI) systems are spreading to evermore sensitive areas of human life. More generally, (less sophisticated) software systems, including decision support systems (DSS), which have been used for decades at this point, influence our lives in countless ways.<sup>1</sup> They are used in autonomous vehicles (Levinson et al., 2011), to support hiring decisions (Langer et al., 2018), to interpret medical images in search of indications of cancer (Kourou et al., 2015), to determine recidivism scores for convicts and help determine their sentences (Hartmann & Wenzelburger, 2021), and so forth. Many of these applications are quite advanced and err less often than humans (McKinney et al., 2020). Their use not only saves their users' time but often also helps to achieve appropriate outcomes and to prevent unwelcome or harmful consequences, e.g., car accidents or wrong medical treatments, even though these systems are not immune to error themselves.

However, many such systems are black boxes: while users can often access the systems' inputs and outputs, they cannot access or understand, let alone reenact, what happens inside the system. One reason for this is that artificial neural networks and other non-linear machine learning systems usually employ models that involve subsymbolic representations such that even developers or data science experts cannot comprehend their inner workings (Bathaei, 2018).<sup>2</sup> This is often taken to be problematic especially in sensitive situations and has several bad consequences for their use: It is difficult to detect errors in the system's operations, and (arguably) neither users nor affected parties can reasonably trust such systems or make well-founded decisions based on a decision support system's recommendation, seeing as they cannot understand what underlies it. Moreover, one may worry that AI systems infringe on users' autonomy (e.g., due to nudging or outright forms of manipulation) if the systems' behavior is not interpretable to their users. And, given that it is impossible to recognize erroneous decisions or misleading recommendations, it may be difficult, and in some cases impossible, to appropriately attribute responsibility and hold anyone accountable, although in many sensitive situations it is *desirable* to be able to hold someone accountable (a claim we aim to support in Sect. 2). This problem of responsibility—the inability or difficulty of holding someone accountable even when doing so is desirable, which will be spelled out in more detail below—is the topic of this paper.<sup>3</sup>

<sup>1</sup> For ease of exposition, we will use “AI system” broadly to refer to software systems generally in the opening sections. Our example cases will focus on decision support systems.

<sup>2</sup> Another reason for this is of practical nature. Often such systems are proprietary and companies are afraid to lose their competitive advantage if they offer insights into the systems' inner workings. As we are interested in the general problem that comes with black boxes, we focus on principled black boxes and leave such practical considerations for policy and regulatory experts.

<sup>3</sup> A prominent position addressing problems of responsibility in this area is the account of meaningful human control developed by Santoni de Sio and collaborators (Santoni de Sio & Van den Hoven, 2018; Mecacci & Santoni de Sio 2020). We are less optimistic than they are about the prospects of ensuring responsibility in the case of fully autonomous systems, and so will argue for the need to keep a human in the loop (in many cases). Further, explainability of AI systems will be a central component of our strategy to deal with responsibility gaps, whereas Santoni de Sio et al. propose that their conditions of tracking and tracing can be met by way of strategies that do without explainability. Another contrast is that they focus on the control condition of moral responsibility, whereas our approach will draw on the epistemic condition. Finally, their account of meaningful human control builds up specifically on Fischer and Ravizza's (1998) notion of guidance control; we take our account to be more easily compatible with a broader range of approaches to moral responsibility understood as a kind of accountability. That said, their account is congenial to our proposal in many ways. Keeping a human in the loop and providing them with reason explanations—as we will propose—may well be *one* way of ensuring meaningful human control.

That the opacity of AI systems gives rise to these problems is intuitively plausible. Arguably, solutions have to put users and people affected by automated or algorithmically supported decisions in a position to understand what underlies the decisions of the systems. In other words, explanations must be provided.<sup>4</sup> The goal of research in explainable AI (XAI), consequently, is to open the black box, or at least to make it more translucent and perspicuous (Langer, Oster, et al., 2021). XAI, understood in a broad sense, is pursued by researchers from a range of disciplines.

This multidisciplinarity comes with a lot of different perspectives and focuses. For instance, a whole host of papers revolve around problems like those mentioned in the previous paragraphs; they provide arguments for XAI from the broader context of morality or society in general (e.g., Asaro, 2015; Binns et al., 2018; Cave et al., 2018; Floridi et al., 2018; Langer, Oster, et al., 2021; Lipton, 2018; Wachter et al., 2017). However, these discussions do not always tell us how *exactly* we can get from a need for reasonable trust, human autonomy, accountability, responsibility, or the like, to a requirement for explainable AI systems. Moreover, they typically do not tell us which kinds of explanations should be given to meet these concerns.

At the same time, there is a broad variety of technically minded papers from computer science introducing and discussing concrete methods for coaxing explanations out of AI systems (e.g., Bach et al., 2015; Kim et al., 2018; Montavon et al., 2017; Ribeiro et al., 2016; Selvaraju et al., 2017). These papers, however, simply presuppose that their results will help to fulfill the proclaimed requirement. This is not surprising, since they are usually not informed regarding the richness of the nature of concepts like *explanations*, *explaining*, *interpretation*, or *understanding* (Miller, 2019; Miller et al., 2017).

Finally, there are a few papers, such as Wachter et al. (2018), Zerilli et al. (2018), and Miller (2019), that strive to provide a more philosophically and psychologically informed picture of the explanations that AI systems should give. However, despite proposing particular kinds of explanations (viz., intentional, counterfactual, or contrastive explanations), they remain silent on whether, or how, explanations of these kinds meet the needs which motivate the call for explainable AI to begin with, such as enabling reasonable trust, human autonomy, or responsibility. To sum up, there is little discussion of whether and how specific forms of explanations—to be provided by technical tools mentioned in the previous paragraph—deliver precisely what the arguments from a societal perspective in favor of XAI demand.

Against this backdrop, our aim is to combine ideas from all three types of papers: We begin by defending and clarifying the claim that there is a desideratum to be able to hold an individual morally responsible for morally problematic AI-supported decisions or actions in Sect. 2. We then argue that such decisions should often be made by a human in the loop who receives recommendations from a decision support system (Sect. 3). Next, by appealing to the epistemic condition on moral responsibility, we substantiate the claim that the outputs of many such decision support

<sup>4</sup> For a general framework for relating the above and other societal desiderata of various stakeholders to explainability, see Langer et al. (2021a). Specifically, for the case of (reasonable) trust and trustworthiness, see Kästner et al. (2021). Furthermore, see Chazette et al. (2021) for a general model of the impact of explainability on various social and technical phenomena.

systems must be explainable for the human in the loop for her to bear responsibility (Sect. 4). By appealing to cases of disagreement between DSS and human in the loop, we argue that explanations of a certain kind—viz., reason explanations—are especially suitable for enabling morally responsible decision-making (Sect. 5). We conclude with some practical challenges for developing reason-giving XAI systems (Sect. 6).

## 2 The Challenge of Adequate Responsibility Attribution

The call for XAI is often motivated by appeal to worries about high-stakes situations<sup>5</sup> in which moral harms may result from opaque systems, among them the worry that missing explainability leads to an inability to hold anyone accountable, or responsible, if something goes wrong. Let us first turn to why exactly it is important to be able to (appropriately) ascribe responsibility when AI systems are operating, and then turn to the question what is needed to be able to do so. For this, we need to understand what is meant by “responsibility” in the relevant contexts, starting with clarifying the concept of responsibility. To do so, we compare and contrast it with the related legal concept of accountability.

Problems of legal accountability are central to the legal concerns with XAI, for instance, in connection with discussions of an alleged EU Right to Explanation (Wachter et al., 2017).<sup>6</sup> Unfortunately, the term “accountability” is used in a variety of ways in this debate.<sup>7</sup> Decision-makers (and agents generally) are accountable, in the sense in which we are interested, when they can appropriately be *held* to account, i.e., when it is appropriate to demand that they explain or justify their conduct or, further, when they deserve reprimand or punishment, given that their decisions or actions are unlawful (Zarsky, 2013; Edwards & Veale, 2017; see Duff, 2007 and 2019 for a nuanced picture of criminal responsibility).

This legal term is structurally quite similar to philosophical notions of moral responsibility (Talbert, 2019).<sup>8</sup> Moral responsibility for an action, as discussed in philosophy, is often spelled out in terms of the agent’s blame—or praiseworthiness for the action, where this is understood in terms of its being fitting to have certain emotions towards the agent such as resentment, indignation, anger,

<sup>5</sup> As a current example, we refer to the proposal of the Artificial Intelligence Act of the European Commission: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>. Accessed September 29, 2021.

<sup>6</sup> See the General Data Protection Regulation (EU) 2016/679 (GDPR), <https://data.europa.eu/eli/reg/2016/679/oj>. Accessed September 29, 2021.

<sup>7</sup> For instance, “accountable” is sometimes used as indicating that persons that are accountable have to be ready to justify their decisions or actions upon request; as more or less synonymous with “explainable” (Kroll et al., 2017); or as concerned with fair and effective governance (de Laat 2018; Perel & Elkin-Koren 2016.). We will put these uses to one side here.

<sup>8</sup> One author who insists on structural similarities between moral and criminal responsibility is Duff (2007; 2019). Duff (2007) discusses the widely held claim that criminal responsibility presupposes moral responsibility. For opposition to this claim, see Shoemaker (2012; 2013).

or gratitude (Strawson, 1962). This approach has been developed in contemporary debates in various forms (see, e.g., Wallace, 1994; Watson, 1996; McKenna, 2012; Pereboom, 2014; Shoemaker, 2015). The corresponding notion of responsibility is often called “accountability,” and is distinguished from other notions of moral responsibility such as attributability or answerability. Though much of what we argue may hold for different forms of responsibility, we are concerned with responsibility primarily in the sense of appropriate praise—or blameworthiness, as exemplified by Shoemaker’s (2015, 113) notion of accountability: “One is an accountable agent just in case one is liable for being a fitting target of a subset of responsibility responses to one – a subset organized around the paradigm sentimental syndrome pair of agential anger/gratitude – in virtue of one’s quality of regard.”<sup>9</sup> In the following, when speaking of responsibility, accountability is what we have in mind.

Moreover, while it may be that moral responsibility presupposes causal—or more broadly—counterfactual responsibility, it goes beyond that concept: It may be that a moral harm would not have come about had I acted differently, but still I am not blameworthy, for example, because I was not aware of what I was doing. Relatedly, computer scientists will say that certain components of a system are accountable or responsible for its failure, i.e., the failure is counterfactually dependent on the performance of these components, but this does not amount to a claim that they are morally responsible or accountable in the sense we are concerned with (Chockler & Halpern, 2004; Halpern & Pearl, 2005).

Here is an everyday example for the kind of moral responsibility we are interested in. Imagine that human resources (HR) manager Herbert, who is tasked with deciding which applicant will get an important management position in the company, disqualifies April, a black female applicant, because of her race and gender. Herbert is not seriously psychologically impaired. It is therefore appropriate to respond to Herbert’s action by blaming or reproaching him for his behavior, and in this case even by taking up legal measures against him for discriminating against the applicant (see, e.g., Title VII of the US Civil Rights act of 1964 or the German *Allgemeine Gleichbehandlungsgesetz*). That is to say, Herbert is morally responsible and legally accountable for his action.

The use of AI systems can challenge the ascription of this kind of responsibility. Suppose Herbert’s company employs a fully automated hiring system to screen, rank, and

<sup>9</sup> As the inclusion of *praiseworthiness* (for one’s good deeds) indicates, moral responsibility is not an exact analogue of *legal accountability*, which is associated with one’s bad deeds only.

We assume that moral responsibility can be ascribed for decisions as well as actions, but also for their outcomes, and that in the cases we discuss, what is at issue is responsibility both for actions and the decisions that underlie these actions, and also their outcomes. Consequently, we speak of moral responsibility for actions and moral responsibility for decisions interchangeably in this paper. Where appropriate, we also speak of responsibility for the outcomes of an action or decision.

Further, agents are morally responsible for actions *under a description*. In the case we focus on in this paper, it may be that Herbert the HR manager is responsible for rejecting April’s application, but not responsible for discriminating against her—in terms of the “action under a description” terminology, he may be responsible for his decision or action under the description “rejecting April’s application,” but not under the description “discriminating against April.” We say more about this below.

select job applicants. Assume that the system ranks April in the last place and excludes her from the further hiring process. Now maybe this ranking was decisively influenced by the fact that April is a Black woman, or some other irrelevant information. If this is the case, this intuitively raises multiple concerns. One is the question of unfair algorithms and algorithmic bias and discrimination (e.g., Garcia, 2016).<sup>10</sup> Another is the worry that no one can be held morally responsible or legally accountable for excluding April, for there was not *anyone* who excluded her. Matthias (2004) calls this a “responsibility gap”.<sup>11</sup> This responsibility gap, understood as an accountability gap, will be the focus of our paper.<sup>12</sup> We will concentrate on cases of responsibility for biased AI-supported decisions since there is much discussion of algorithmic bias.<sup>13</sup>

Let us sketch two motivations for closing the moral responsibility gap, for making sure that there is a person who can be properly held responsible for such morally problematic decisions. We do so by focusing on the case of Herbert and April. On the one hand, there is a motivation from incentives: If someone like Herbert is morally responsible for the problematic decision or action, this means that he can fittingly be blamed for it. It is then, at least *pro tanto*, just to express blame or even to establish legal sanctions (McKenna, 2012, though there may be exceptions). This will plausibly motivate him to be diligent in making up his mind about whether to (follow the system’s recommendation and) disqualify the black female applicant to avoid negative consequences for himself. Such an incentive for diligent decision-making may lead to better hiring decisions and less wrongdoing (for empirical evidence, see Fehr & Gächter, 2002).<sup>14</sup>

<sup>10</sup> We cannot enter into the debate about discriminatory and unfair models here. Roughly speaking, a model is unfairly biased against members of a certain group if (and only if) it treats its members quantitatively unequally (to their disadvantage and without justification), for instance, by making certain classificatory errors more or less frequently. This particular unequal treatment may either have immediate causes in input data, for example, when the system directly refers to April’s characteristic of being Black, or it may be caused indirectly, by giving certain proxy variables undue weight, such as April’s Alma Mater, her hobbies, or her zip code. A range of technical issues may be behind such biases in systems, e.g., when sensors perform worse for darker skin colors. Furthermore, there may be problems with the quality of training data, for example, when the data does not adequately represent the population or when it is a result of a biased social process. Some unfair treatment is generally mathematically unavoidable (Chouldechova 2017; Lepri et al., 2018).

<sup>11</sup> For overviews, see, e.g., Johnson (2015) or Noorman (2020).

<sup>12</sup> The terminology in the responsibility and responsibility gap debates is not unified. For instance, our accountability gap corresponds largely to Santoni de Sio and Mecacci’s (2021) culpability gap, but not to their accountability gaps.

<sup>13</sup> Related worries about moral responsibility can be raised by other cases: For instance, one might worry about how properly to allocate responsibility in case a system simply does not work reliably. Furthermore, there may be a DSS that recommends a decision that is optimal given only its available information. However, the human decision-maker has further information available that, put together with the system’s information, shows that this decision is terrible. Yet without access to the information that the system relied on, the decision-maker is unable to put two and two together, so to speak, and therefore unable to figure out what would be the best decision overall in the situation. It seems wrong to hold the decision-maker responsible for making a bad decision in this example. We discuss a case of this kind in Sect. 5.

<sup>14</sup> While (possibly unjust) expression of blame and punishment could of course also incentivize agents who cannot be held responsible, such incentives do not seem palatable to many, and going this far is not necessary as long as someone can be held accountable to whom incentives can be more appropriately applied. Furthermore, an agent who meets a central condition of responsibility is better placed to comply with incentives to avoid harming others (as we argue in Sect. 4).

We acknowledge that this argument needs further details in order to evade counterarguments. For instance, it has been put to us that one may always be able to find someone responsible for producing or employing the DSS if it discriminates against applicants, and that person will have an incentive to be diligent that no discrimination arises. However, as we point out in the next section, if the system's discrimination is not *foreseeable* to anyone, there may be no one bearing indirect responsibility of this kind. Furthermore, it might be that only someone at the company who developed the system can be held indirectly responsible but nobody at the companies that employ the system. Then, the system might be applied carelessly by a great number of users who need not bother as long as, for instance, the system is not taken off the market. In this case, there would be no incentive for applicants to avoid wrongdoing in hiring decisions.

On the other hand, there is a motivation of justice: If April suspects—or finds out—that she was discriminated against because of her race and gender, it would intuitively be desirable to enable her to blame someone for wronging her. It would be desirable to make it possible for her to get justice, in the sense of a *person responsible* for discriminating against her owning up to the fact that they did something wrong. She should be able to be fittingly angry with someone and to express this anger by demanding of a responsible decision-maker that they acknowledge their wrongdoing, that they apologize, make amends; it would be desirable to make it possible that they get sanctioned. To motivate this further, imagine that the responsibility gap cannot be closed. Then April's situation is morally equivalent to the situation of another agent, call her Berta, who has been harmed by a natural disaster: Both April and Beth are harmed, nobody is responsible, and nobody is blameworthy. However, April's and Berta's situations are intuitively different. Many people were involved in setting up and using the system that harms April, but no human is involved in harming Berta. And it seems that this makes a difference in terms of justice, for Berta really cannot justly blame anyone, but intuitively April *should* be able to appropriately blame someone and may reasonably desire to do so.

Of course, here too one may raise doubts, for instance, by questioning whether justice requires being able to angrily blame someone or just being able to do something in the vicinity. Maybe all that is needed is someone who is *answerable* in the sense explicated in Shoemaker (2015, 82),<sup>15</sup> i.e., someone who is able to cite their reasons for the action and who is thereby liable for being a fitting target of responses like agential regret or pride in virtue of their quality of judgment. In our concrete case, this person would be expected to admit and regret a discriminatory hiring decision. Such answerability would not imply accountability.

Arguably, that there is someone who is answerable might already be helpful to some degree to ensure justice for the wronged applicant. However, in the case of discrimination and other offensive treatment, it would further be desirable for an agent like April to be able to fully hold someone accountable. While it seems

<sup>15</sup> Further explications of answerability that do not imply accountability can be found in Smith (2005) and Scanlon (2008, chap. 4). Duff (2007) takes answerability to consist, roughly, in someone's being an appropriate target of the request for an explanation. We thank an anonymous referee for pressing us to address the counterexamples to this and the following arguments.

right that April deserves an explanation, she should also be able to be fittingly resentful for being disadvantaged based on her race and gender, and to be able to call for moral sanctions in terms of blame. This indicates that accountability and not just answerability is relevant (see Shoemaker, 2011, 616 and 621).<sup>16</sup> We acknowledge that these questions can be debated further. However, since our main focus is an argument to the effect that explanations are often the best way of closing the moral responsibility gap, it is sufficient for our purposes here to present these initial motivations that could be spelled out further. In our view, the argument from incentives and the justice-based argument provide a compelling rationale for a desideratum to avoid high-stakes situations in which no one can be held responsible.

### 3 Why We Need Someone in the Loop

But how to make sure that there is a person who can properly be held responsible? In this section, we will argue that, if we want to ensure that a human can bear responsibility for morally problematic decisions, we often cannot—and, in fact, should not—delegate these to fully automated systems. Instead, we should keep a human in the loop: AI systems should be used merely to supply recommendations about what to do, but the final decision should be left to a human decision-maker—in our example, to Herbert.<sup>17</sup> However, keeping a human in the loop is, as we argue in the next section, not sufficient to ensure that there is someone who can bear responsibility. But before we can turn to the question of what is missing for a sufficient condition—and how this relates to explainability of a certain kind—we want first to give a convincing argument for requiring a human in the loop.

The obvious alternative to keeping a human in the loop to bear responsibility would be to find someone else at the company, someone who decided to purchase the system or (one of) the developers of the system and to allocate moral responsibility for the specific fully automated decision to that person. What speaks against this alternative? In a fully automated decision process, no one made the decision or was able to influence it directly. So no person can bear *direct* responsibility for (the outcome of) the fully automated decision. A person at the company or a developer could, at most, bear responsibility *indirectly* given that the decision was fully automated. Indirect responsibility can be ascribed to an agent for an outcome where she is directly responsible for something else—such as

<sup>16</sup> The same argument could be phrased in a more victim-centered way: Plausibly, in a situation in which applicants are systematically discriminated against, many of them will *desire* to be able to at least hold decision-makers or their companies accountable for discriminatory decisions, to be able to resent and blame them. This is a good reason to ensure that someone can indeed be held responsible.

<sup>17</sup> Human in the loop cases contrast with cases where a human is on the loop: Here, a human supervisor is informed about the decisions of an AI system before they are put into effect and can step in if need be. We will focus on the “in the loop” scenario for sake of simplicity. What we say for this scenario can be modified and applied to many, though not all “on the loop” scenarios.

her own ignorance or loss of control—which led to that outcome. In such cases, this “something else” is her fault, for instance, because she did not do enough to meet obligations to stay informed or in control (Mele, 2021; Rosen, 2003; Zimmerman, 1997).<sup>18</sup> One might think that someone will bear indirect responsibility for the fully automated decision by being responsible directly for employing—or designing—a faulty system, so we can also blame them, indirectly, for the particular decision made by the system.

The proposed assignment of indirect responsibility, however, runs up against an especially nasty variant of the problem of many hands (Thompson, 1980; van de Poel et al., 2015). The problem of many hands, as we understand it here, is quite generally that, in a complex situation, in which the contributions of many agents lead to moral harm, such as when large corporations and companies cause a problem, it is difficult to allocate direct or indirect moral responsibility to anyone in particular. The problem of many hands has an epistemic and a metaphysical dimension: On the one hand, it is concerned with difficulties in determining who is morally responsible and, on the other hand, with difficulties with respect to whether anyone actually is responsible.<sup>19</sup> It further has a practical-political dimension: Even if there is someone who is—directly or indirectly—responsible, complex situations with many contributors lend themselves to obfuscation, making it easy for companies and other agents to let themselves off the hook.

Adding a fully automated AI system to the mix compounds the problem. Suppose that a level 5 self-driving vehicle<sup>20</sup> kills a pedestrian. Should the blame and thus the moral responsibility for the accident be allocated to the company who built the car, to the company who supplied the car’s LiDAR (light detection and ranging), or to the company who owned and employed the vehicle for the mission it was undertaking, etc.? If one of the companies is held responsible, which person at the company is to bear responsibility? It appears that an already complex situation here is made even more confusing by the involvement of an autonomous AI system (Awad et al., 2018; Coeckelbergh, 2020; de Laat, 2018; Mittelstadt et al., 2016; Nissenbaum, 1996; Sparrow, 2007). Having a human in (or at least on) the loop, by contrast, alleviates the problem of many hands by providing at least one easily detectable and plausible candidate for bearing the direct responsibility for a particular decision that caused harm. After all, if the car were operated by a human who could reject any recommendation or decision of the system, that person is an

<sup>18</sup> A related notion is that of “tracing,” which was introduced by Fischer and Ravizza (1998, 49).

<sup>19</sup> Problems of many hands can arise in virtue of independent decisions of several different agents, which are each sufficient for a harmful result. Since the result is overdetermined, no one agent could have prevented it by acting differently, so it seems each agent is off the hook. Such problems can also arise when each contribution of several agents by itself would have been harmless, but all actions in combination lead to a morally problematic result. Here, it may be that no agent by herself could have prevented the result, so again it is difficult to ascribe moral responsibility to anyone.

<sup>20</sup> See SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, [https://www.sae.org/standards/content/j3016\\_202104/](https://www.sae.org/standards/content/j3016_202104/). Accessed June 23, 2021.

obvious candidate for blame.<sup>21</sup> This is not to deny that it may be important to allocate additional indirect responsibility to the companies involved if these could and should have done more to prevent accidents.<sup>22</sup>

One central way in which the problem of many hands may arise is that, because of the many agents involved in a situation, agents contributing to a harm are unable to foresee that their combined actions will lead to a problematic outcome. We will explore the role of knowledge for responsibility in more detail in the next section, but even pre-theoretically, it seems problematic to hold someone responsible for a harmful outcome when it was not foreseeable by them, and so when they were not at fault for not foreseeing it. Transferring this worry to the case of fully autonomous AI systems like the hiring system used by Herbert's company, it may well be the case that all people who might bear moral responsibility indirectly for the output of an AI system blamelessly lack relevant (fore-)knowledge regarding the system's output. This may be true, e.g., of the developers of an AI system, of people at an accreditation agency, and of the companies employing the system. Even a thoroughly tested and generally reliable system may give a problematic output when certain features of the situation to which it responds combine in an unusual way (Edwards & Veale, 2017).<sup>23</sup> Focusing on the issue of bias, systems cannot easily be tested before or during employment with respect to bias because the biases for which they would be tested concern protected classes, such as sexual orientation, which often is not available in the data to the developers (Lepri et al., 2018). Furthermore, bias might be hidden in the statistics—even though a system, say, puts women at a disadvantage as compared to men, the overall statistics regarding its performance may look unproblematic. Thus, there are likely many cases in which neither developers nor the persons employing the system can be expected to foresee particular harmful future outputs of a system. And so they cannot properly be held (indirectly) responsible for them.

A final problem for allocating indirect responsibility is that there may be cases in which, although we may succeed in finding a person (e.g., developers or employers of the AI system) to whom we can ascribe indirect responsibility for a particular output, it would be wrong to hold them responsible nonetheless. For it could be that the system is the best system that could have been developed—e.g., it is the least biased hiring system that it is possible to design—but still there are some fringe cases in which its output is biased, e.g., in that it puts Black queer women from a low socioeconomic background at a disadvantage. To put it differently, it may

<sup>21</sup> In the Uber car accident in Tempe, Arizona, in March 2018, a self-driving vehicle owned by Uber was operated by a human on the loop who was supposed to intervene in case of emergency; otherwise, the car was in self-drive mode. Here, the operator is the most salient candidate for bearing responsibility, although not the only one. See Will Pavia (March 21, 2018), Driverless Uber car "not to blame" for woman's death, *The Times*, <https://www.thetimes.co.uk/article/driverless-uber-car-not-to-blame-for-woman-s-death-klkb7vf0>. Accessed June 23, 2021.

<sup>22</sup> A human in the loop is ideally positioned to realize that a DSS is not working properly, e.g., by exhibiting bias in many recommendations. She is thereby able, in principle, to recognize that the developers of a system are at fault for creating a malfunctioning system, for which they can be held morally responsible.

<sup>23</sup> This concern relates in interesting ways to the issue of intersectionality discussed in feminist theory: The experience of discrimination made, e.g., by a black woman, is different and potentially worse than that of a white woman or a black man, since different categories/dimensions of discrimination combine (Cooper 2016).

be that the overall great performance can only be achieved at the price of allowing some suboptimal outputs in rare cases. We can even imagine that the system is much better overall than a human decision-maker would be. If so, it seems that the developers or employers of the system have done nothing wrong, and so cannot be blamed. Nonetheless, it would be desirable to have someone who can bear responsibility for the individual biased hiring decision and its morally problematic outcome.

Even if a candidate for indirect responsibility could be identified (contrary to the epistemic dimension of the problem of many hands), the unpredictability of problematic outputs and the issue of overall optimal performance may prevent that candidate from bearing responsibility. This holds both for the developer of the AI system and for a customer who relies on an accredited system that she leaves to operate by itself. Bearing moral responsibility for a particular output of an AI system, even indirectly, requires that there is someone who is able to foresee it and who cannot evade blame because they did the best they could. As argued, these conditions will often not be met.

With this argument in place, we need to add a qualification. Fully automated or autonomous AI systems may be acceptable in some cases. For instance, two of the most pressing and widely debated applications discussed in the context of responsibility gaps are autonomous driving and (lethal) autonomous weapons. Autonomous driving typically involves no human at all or at most a human on the loop. Similarly, while the mode of operation for drones in general has been moving more and more from human in the loop to human on the loop setups,<sup>24</sup> lethal autonomous drones involve at best a human on the loop, who can interfere with the decisions of some autonomous system that identifies potential targets. In both cases, the time available to make a decision may not be sufficient for an effective handover, let alone an “explained handover,” even if such a handover is technically possible. In light of this, a human in the loop and concurrent explainability of an output to this human may not be all things considered the best way to go, even if this entails *pro tanto* undesirable responsibility gaps. For example, assume that autonomous vehicles<sup>25</sup> prove to be clearly superior to human drivers in certain contexts, so that critical situations only occur in a fraction of cases, while the time for an (explained) handover is too short. If this is the case, it may be that a fully automated set-up is in some cases significantly better than one involving a human in the loop, so that it *may* be all things considered permissible to leave corresponding responsibility gaps open.<sup>26</sup> However, we believe that this is true only of a limited number of cases (e.g., some cases with extreme time pressure or very low stakes), so that our argument gets a grip in a significant number of cases.

<sup>24</sup> See United States Air Force Unmanned Aircraft Systems Flight Plan (May 18, 2009): [https://irp.fas.org/program/collect/uas\\_2009.pdf](https://irp.fas.org/program/collect/uas_2009.pdf). Accessed September 29, 2021.

<sup>25</sup> Let us make explicit that we make the following argument for autonomous vehicles, but *not* for lethal autonomous weapons systems. Here, we believe that the stakes are too high for the advantages of a fully automated setup to outweigh the disadvantages of not being able to hold anyone directly responsible.

<sup>26</sup> Similarly, responsibility gaps may be acceptable in low-stakes decision situations, e.g., involving movie recommendation services on streaming platforms. While it may be *prima facie* desirable to have a responsible agent for every single recommendation (as they may be discriminatory as well), plausibly the stakes are simply too low to justify the effort.

Next, are we not letting developers and companies employing decision support systems off the hook too easily? This is not so. Note the following two features of our argument. First, the strength of our claim: Our aim is to establish that keeping a human in the loop (and providing them with explanations) is *one* good way of ensuring that we can properly hold someone responsible. We suggest that, in some contexts, this may be the best or even the only way to go, but leave open that there may be other ways for ensuring responsibility more suitable for other situations (and some of these ways may rely on explainability or other perspicuity enhancing capabilities *after* the fact, to determine what went wrong in the relevant situation, see Sterz et al., 2021).

Second, the scope of our claim: Our focus is on moral responsibility, and how to ensure that there is an agent who can properly bear it in the context of AI-supported decision-making. Whether the same argument applies to related phenomena such as legal accountability is a further issue beyond the scope of this paper. One suggestion is that, even if it is not possible properly to ascribe moral responsibility to the developers or employers of an AI system, we may still be able to hold them accountable by law (e.g., by imposing a strict liability for damages arising from the operation of a car on its registered keeper). The current debate over the German law for regulating automated driving, which has been criticized for making vehicle owners liable for damages instead of manufacturers, indicates that similar problems arise in legal contexts.<sup>27</sup> Finally, we allow that there may be cases in which it is justifiable not to enable moral responsibility, e.g., where affected parties are compensated for not having someone to hold responsible.

Overall, we conclude that the allocation of indirect responsibility is often infeasible. Instead, we then need a person who is presented with the output during use and has the chance to interfere—a human in the loop. Since a human in the loop is made knowledgeable of the recommendation during use and makes the relevant decision herself, she is a candidate for direct responsibility for the outcome.<sup>28</sup>

## 4 Connecting Responsibility to Explainability

A human in the loop is a candidate for responsibility, but there are further requirements to properly allocate responsibility to them. This is where the demand for explainability comes in. As Floridi et al. put it, ensuring “that the technology – or, more accurately, the people and organizations developing and deploying it – are held accountable in the event of a negative outcome, … would require … some *understanding* of why this outcome arose” (2018, p. 700, our italics).<sup>29</sup> To have such understanding, the human in the loop, at the time of the decision, needs access to

<sup>27</sup> Gerald Traufetter (May 7, 2021), Vorstoß von Verbraucherschützern – Hersteller sollen bei Unfällen mit autonomen Fahrzeugen haften, *Der Spiegel*, <https://www.spiegel.de/auto/autonomes-fahren-hersteller-sollen-bei-unfaellen-haften-fordern-verbraucherschuetzer-a-78df0b3a-0002-0001-0000-000177426963>. Accessed June 22, 2021.

<sup>28</sup> In the following, when speaking of responsibility, we refer to direct moral responsibility, if not stated otherwise.

<sup>29</sup> For a useful discussion of this issue in the medical context, see Robert David Hart (September 10, 2018), “Who’s to blame when a machine botches your surgery?” <https://qz.com/1367206/whos-to-blame-when-a-machine-botches-your-surgery/>. Accessed June 23, 2021.

an explanation of the DSS's recommendation and possibly its overall functioning. Our aim in this section and the following is to motivate and substantiate the claim that explainability is needed to make the human-in-the-loop solution work, and to investigate what kind of explainability would do the job well. We do so by focusing on what a human in the loop needs in order to meet a standard condition for moral responsibility: This is, at bottom, an explanation of the system's output.

The foundation for our reasoning lies in a necessary condition on direct moral responsibility which is widely discussed in the philosophical debate—the *epistemic condition* (Noorman, 2020; Rudy-Hiller, 2018). According to it, an agent is not directly morally responsible for an action unless she is aware, or in a position to be aware, of what she is doing, of the (probable) consequences of her action, of its moral significance, or of alternative options available to her. For instance, an agent who flips the switch to turn on the light and who thereby electrocutes her neighbor by an unfortunate combination of circumstances that was not foreseeable is not directly responsible for the harm caused.

One way to make this distinction more tangible is to resort to a coarse-grained view of actions according to which one action can be picked out under a range of different descriptions (Anscombe, 1962; Davidson, 1963). In the example, the agent's action can be described as flicking the switch *or* as turning on the light *or* as electrocuting the neighbor. Since she is not in a position to be aware that her action is one of electrocuting her neighbor, she is not directly morally responsible for it under that description, though she may still be responsible for flicking the switch. For Herbert and April, the crucial question then, in the context of the epistemic condition, is not whether Herbert is responsible for rejecting April, but whether he is responsible for discriminating against April.<sup>30</sup> In light of this distinction, the epistemic condition can then be spelled out thus:

(Epistemic Condition) An agent is morally responsible for her action or decision only if she has sufficient epistemic access to it. *That she has sufficient epistemic access to it entails at least that she is in a position to know the action under relevant descriptions.*<sup>31</sup>

The epistemic condition on moral responsibility can be used to provide two motivations for making decision support systems explainable—the first motivation will be introduced by appealing to an initial case, and the second by appealing to a fleshed-out version of this case. Our *initial* case is the hiring case in which HR manager Herbert is a human in the loop and makes the final hiring decision, but does not have an explanation of the hiring system's recommendation. Assume that, before his company started to employ the decision support system, Herbert used to be the HR manager who competently and responsibly made hiring decisions for his company,

<sup>30</sup> We will leave out the qualification ‘under a certain description’ in the following, except where it is necessary for our argument. We assume that the relevant description is clear from context. Plausibly, the agent's role (such as being an HR manager) helps to determine the descriptions under which we want to hold her responsible for her action or decision.

<sup>31</sup> Note that the question of what descriptions exactly are relevant goes beyond the scope of this paper. This determination might well be a function of the specific context of the decision-situation. For an idea of how to operationalize context-sensitive societal desiderata more generally, see (Köhl et al., 2019; Langer et al., 2021a).

and that he will continue to do so, using the DSS's output as one source of support. We focus on human in the loop cases like Herbert's, in which the decision-maker relies on a DSS and no other AI systems play a role.

Imagine that Herbert decides to exclude April's application because the hiring system recommended doing so. Imagine further that the system's recommendation is due to its bias against Black female applicants, but that, since it is an accredited system, Herbert justifiably believes that it has no such problems. Herbert is therefore not indirectly responsible for discriminating against April—he is not to blame for being unaware of the system's bias. If he is responsible, he must be directly responsible, which requires his being in a position to know what it is that he is doing, its probable consequences, and its moral significance. As described, if he does not have access to what moved the DSS to provide its recommendation, then his AI-supported decision will be made without him being in a position to know these things. Herbert is aware that he rejects April's application, and so he is aware of his action under that description. But he is not in a position to know that what he is doing, under another description, is to discriminate against her. Nor is he in a position to know that he unfairly rejects her application and that this is an act of moral wrongdoing. Consequently, he is not morally responsible for discriminating against April.

Once a meaningful explanation of the recommendation is available to the decision-maker, we can more easily bridge the responsibility gap. For instance, assuming that the system discriminates against April immediately based on her race and gender, then, if Herbert has access to this fact, he does have access to—is in a position to know—the fact that to reject her application on this basis is to discriminate against her; and that it is unfair and an act of moral wrongdoing. But even in the case where the system discriminates against April based on a learned correlation involving some otherwise innocent proxy variables such as, say, April's Alma Mater, her hobbies, and her zip code, explanations may enable Herbert to get the right kind of epistemic access. For the proxies will typically be either suspicious or seemingly irrelevant. In both cases, Herbert should doubt the system's recommendation: If the system indicates that it considers the combination of April's Alma Mater, her hobbies, and her current zip code to be particularly crucial, this may catch Herbert's attention: Is this not one of the historically Black colleges and universities? And is that not a primarily Black neighborhood?

In any case, an explanation allows Herbert to become suspicious and to pay particular attention to the role played by other factors. Herbert can then check, if necessary, whether candidates with otherwise similar profiles are rated similarly. In this case of proxy-based discrimination, Herbert may not be sure that discrimination is present, but given sufficient background knowledge and awareness of the danger of discrimination by models, he can develop an initial distrust and at least begin to consider that other descriptions of the situation might be relevant. He is therefore in a position to know at least that a decision that follows the system's recommendation *may very well be discriminatory*. So, while explanations may not guarantee in all cases that the epistemic condition on moral responsibility is met, they clearly facilitate its fulfillment.<sup>32</sup>

<sup>32</sup> If the required background knowledge of the danger of discrimination by models is unavailable to Herbert, he may fail to be responsible for discriminating against April, since the proxies may then not raise his suspicions. But then maybe someone else is (indirectly) responsible for not having ensured that Herbert has what it takes to fulfill his role competently.

Let us turn to our *second* motivation. At least on one way of fleshing out Herbert's situation further, his epistemic situation is even worse than has become apparent so far. Our fleshed-out scenario shows that, if a decision-maker cannot tell why a DSS provided the recommendation it did, then there may be situations, particularly situations of disagreement between system and decision-maker, in which he cannot tell whether his decisions bring him closer to his goals. As a consequence, he is unable to guide his decisions so as to pursue these goals, or to execute his intentions in acting. This gives rise to an especially threatening way in which an agent lacks epistemic access to his action, and thereby also lacks moral responsibility for it.

Here is the fleshed-out scenario. Imagine that Herbert, at the end of a lengthy selection procedure, is presented with a list of three applicants that the DSS ranks as the top candidates; the system recommends keeping them in the running for the position. April did not make the list, but made it into the top ten. However, Herbert, by going through the top ten applications independently, counted her among the top three applicants beforehand. So we have a case of disagreement between the system's recommendation and Herbert's initial judgment. Since there is no explanation of the system's recommendation available, Herbert cannot reasonably resolve the disagreement.

Here is how this might happen: Say that his own assessment of April's qualities is due to good, but not conclusive reasons—she has more relevant work experience than most; received great grades in her studies at Yale; speaks a foreign language, which is useful but not absolutely necessary for the job; and has work experience abroad. (By saying that his reasons are not conclusive, we mean that they are weak enough that he may reasonably question his own judgment if the system gives a contrary recommendation.) On the other hand, the system was accredited to be reliable by a trustworthy watchdog organization, though Herbert is aware that systems of this kind may have hidden bugs or biases. In this situation, the system's countervailing recommendation leaves open both the possibility that Herbert correctly assesses the situation and the system is mistaken *and* the possibility that the system has a superior understanding of the situation, and Herbert is in the wrong. In the first possibility, the system's recommendation may be due to some kind of bug, or to its bias against women of color; in the second possibility, the system's recommendation may be due to the fact that it has access to information Herbert does not have, or detects patterns that Herbert misses. Say that the system relies on all of Herbert's reasons for taking April to be among the top three candidates (her great grades from Yale, her foreign language competences, etc.). However, it has detected that applicants with these qualifications *taken together* tend to move on to other, better jobs very quickly. So the system detects a pattern which turns what would otherwise be great reasons for hiring a candidate into a reason against hiring her.

This illustrates that, in a particular situation, Herbert may be unable to tell whether he is in one of two relevant cases:

*Case 1* The system's recommendation is mistaken and Herbert's assessment is right.

*Case 2* The system's recommendation is correct and Herbert's assessment is wrong.

Given that the two cases are indistinguishable to him, he cannot reasonably resolve the disagreement. For he cannot compare or reconcile his own and the system's reasons for or against keeping April in the running, and so cannot figure out which reasons are superior, e.g., by weighing them against each other. Consequently, if he decides to keep her in the running, this decision is arbitrary; but if he decides

to exclude her from the short list, that decision is also arbitrary.<sup>33</sup> The lack of access to the system's reasons undermines Herbert's ability to come to a well-founded all-things-considered judgment about which applicants to keep in the running.

In light of his inability to come to a well-founded all-things-considered decision, Herbert is then unable to competently pursue his goal. Say he is genuinely trying to find the best candidate for this prestigious, responsible position at his company. Since he is unable to tell which is the proper means to doing so—keeping April in the running or excluding her—he is thereby unable to respond to pertinent reasons in pursuit of his goal. In other words, he cannot properly guide his decisions in light of his goals, so as to execute his intentions. This undermines his ability to find the best candidate or to reach various related goals. Imagine Herbert instead trying to damage the company by hiring an unsuitable candidate. Again, since he cannot tell whether it is his or the system's assessment of April that is right, he is unable to tell whether excluding April would be a good means to pursuing this goal, and this undermines his ability to guide his hiring decision in response to pertinent reasons.

In the fleshed-out scenario, Herbert is especially epistemically impaired: He is not in a position to know either of his options under the relevant descriptions. He cannot tell whether, if he complies with the system's recommendation, his decision is one that wrongs April; but neither can he tell whether, if he goes with his own initial assessment, his decision can be described as one of harming his company. In this fleshed-out version of the scenario, then, Herbert's access to his decision is undermined in a more severe way. Because of this more wide-ranging epistemic disconnect, Herbert is not directly morally responsible for his AI-supported decision.<sup>34</sup>

<sup>33</sup> But can a human in the loop not decide non-arbitrarily by relying on the system's (or his own) past track record? There may be situations where this is possible, and for these, we concede that the problem is mitigated. However, this leaves many situations in which both the human in the loop and the DSS have equally good or bad track records, where this information does not help. Moreover, it can be hard to determine a track record: For some tasks, such as finding the best applicant out of a pool of applicants, it may be difficult to figure out whether they have been performed successfully, and so also whether a subject or a system has a good track record with respect to the task. We thank an anonymous reviewer for posing this question.

<sup>34</sup> One might use this line of thought to raise trouble for responsibility in three further ways, which we cannot develop here, but would like to at least mention: (1) One might connect it to another widely discussed necessary condition for moral responsibility, the control condition (Talbert 2019), if control is understood to require reason responsiveness. Our case of disagreement arguably undermines the decision-maker's ability to recognize a certain class of reasons, those which lead to the recommendation of the DSS. One could spell this out by appealing to a constraint included in Fischer and Ravizza's (1998, 64) account, viz., that the agent has to be able to act for the sufficient reason that favors his action. If the agent is unable to access the system's reasons, then he is unable to act for them (Mantel 2018).

(2) It might be argued that our case undermines Herbert's (backward-looking) responsibility via undermining his ability to meet his forward-looking responsibility as an HR manager. Plausibly, in virtue of his professional position, Herbert has the task—and therefore the obligation—of finding the best candidates for jobs at his company, without unfairly relying on, e.g., candidates' membership in certain groups. If he fails to meet this professional obligation, he can appropriately be held accountable for this failure. In cases of disagreement, Herbert's inability to know what decisions he has available under relevant descriptions and his consequent inability to competently pursue his goals, undermine his ability to fill his professional role, and so he lacks forward-looking responsibility. So, given that backward-looking responsibility hinges on forward-looking responsibility (Duff 2019), Herbert cannot be held responsible for his hiring decisions in these cases either.

(3) One might explore whether, on some Strawsonian picture, it could be said that, for the agent to be morally responsible, their quality of will must be *expressed* in the action in a way it cannot be expressed in Herbert's case without having access to the system's reasons.

Of course one might object that cases of disagreement are insignificant outliers. Typically, the decision-maker will agree with the system's recommendation. However, this objection renders the use of decision support systems obsolete. If the system's recommendation allows for well-founded decision-making only where it supports what the decision-maker would choose anyway, then it is pointless to combine a DSS with a human in the loop for the hiring decision. From Herbert's perspective, adding the DSS does not improve his decision-making; from the perspective of the company, keeping a human in the loop does not add an advantage over employing a fully automated system. The system can lead to better decision-making exactly by way of disagreement with the decision-maker where there is room for changing his mind. So, *exactly when it counts*—when the decision-maker has reasons that are not conclusive, and the system makes a recommendation that is potentially better than his take on the situation—the system undermines the decision-maker's epistemic access to his decision, and thus his moral responsibility.

Without explainability, we face a dilemma for human in the loop scenarios: It is *either* pointless to have the system provide a recommendation to the human decision-maker (in cases where human and system agree, or when the decision-maker has conclusive reasons anyway), *or* the lack of explainability undermines his epistemic access to his decision and thus the moral responsibility which the human in the loop is supposed to bear (in cases where human and system disagree, while the human has non-conclusive reasons). Now the second horn of the dilemma is due to the fact that the decision-maker has no access to why the DSS provided a certain recommendation. If he had a suitable explanation of the system's recommendation available, so that he would be able to compare his reasons with the system's reasons, he would be in a better position to figure out whether it is the system's or his own assessment of the situation that is correct. So, he would be able to resolve the disagreement in a non-arbitrary way, thereby be able to make the hiring decision that best suits his goal (finding the right person for the job), and thus be in a position to know his decisions and actions under the relevant descriptions. We conclude that, in many cases of disagreement where the decision-maker's reasons are non-conclusive, he is in a position to bear direct responsibility for his decision just in case he has a suitable explanation of the system's recommendation available.

To sum up, a human decision-maker needs explanations. These enable responsible AI-supported decision-making by enabling the agent to meet the epistemic condition in cases like the ones discussed in this section.

## 5 The Advantages of Reason Explanations

Which form should an explanation take to ensure that decision-makers are morally responsible for their AI-supported decisions? While different kinds of explanations could enable responsibility when properly interpreted by human decision-makers, reason explanations are particularly well-suited for this job. They are the ones that humans typically use when trying to understand and explain action, when exchanging justifications for actions and recommendations, and when trying to resolve disagreements (Alvarez, 2010; Hieronymi, 2011). Just like human experts would

provide reasons for their recommendations, so should decision support systems. In this section, we spell out how reason explanations help to resolve different kinds of disagreement between humans in the loop and DSS and what kind of reasons are needed for the job.

Before returning to the disagreement case and illuminating what reason explanations for decision support systems should look like, let us first clarify what reasons are and which kinds of reasons figure in reason explanations. In the philosophy of action, reasons are categorized by the distinction between normative and motivating reasons (Alvarez, 2017; Hieronymi, 2011; Mantel, 2018). We here apply—without defending it—this widely accepted philosophical distinction to the recommendations of decision support systems. *Normative* reasons are facts that objectively favor or disfavor an action (such as the action recommended by a DSS). All normative reasons, taken together, make the action right or wrong. For instance, the fact that eating vegetables is healthy counts in favor of my eating vegetables. Applied to decision support systems, we may say that normative reasons are the facts which favor or disfavor a DSS's recommendation and the recommended action. When a system's input data contains information that fits the facts and supports the recommended action over another, we can say that the system has available normative reasons favoring a certain recommendation.

Although ideally a DSS has normative reasons available, reason explanations should focus on *motivating* reasons instead, because systems can make mistakes. A motivating reason is a consideration that an agent relies on in acting, a consideration “for which someone does something, a reason that, in the agent's eyes, counts in favor of her acting in a certain way”—whether or not it is a fact and actually favors the action (Alvarez, 2017). Motivating reasons stand at the intersection between explanation and justification insofar as they help to explain the output in the light of what the decider took to justify or favor it (Hieronymi, 2011). Unlike normative reasons, motivating reasons can include merely apparent facts, i.e., non-obtaining states of affairs or false propositions that the agent falsely takes to obtain (Dancy, 2000; Schmidt, 2018). For instance, that spinach is a good source of iron is a merely apparent fact. Even though it is not the case that spinach is a good source of iron, this can be the reason which motivates me to eat spinach—since I mistakenly believe that spinach is a good source of iron, in my eyes, this favors the action, and it is the light in which I act. If a motivating reason is not mistaken, we say that it corresponds to a normative reason.

We suggest what one might call a functional picture of motivating reasons, on which “favoring in an agent's eyes” is not interpreted as entailing awareness. We talk of motivating reasons more loosely to pick out information which plays a certain role in determining the output of a system, e.g., in whether or not it recommends a certain action. With this functional characterization in mind, it becomes feasible to transfer reasons to decision support systems. A DSS can then be described as providing recommendations on the basis of reasons available to it, or, to put it differently, as treating something within its inputs as reasons for its recommendation. For it can be correct that the system provides a certain recommendation because it has certain information (i.e., motivating reasons) available. Note that this does not

yet commit us to the claim that there is a form of (non-deflationary) reasoning to be found within that system.

Turning next to *reason explanations*, a reason explanation explains an action in terms of an agent's motivating reasons—that is in terms of the information or misinformation that led her to the action. Ideally then, a reason explanation of a system's recommendation will include only the information on which the system relied in producing its output—the information contained in the data available to the system on which it relied in providing its recommendation. The explanation refers to the information which *actually* contributed to the system's coming to a particular recommendation, and not to confabulations. This is not to say, however, that the reason explanation refers to all the information that made a contribution to the recommendation or decision. Although agents may be aware of a huge number of *pro* and *contra* considerations and may be led to an action by such a bundle of reasons, most reason explanations of human action focus on just one or a few contextually relevant motivating reasons. Even if a DSS takes into account much more information than a human would in providing a recommendation, this complexity therefore does not rule out providing a simple reason explanation for its recommendation, for such explanations typically do not require to name all of the motivating reasons but only the most relevant ones. What it does require, of course, is singling out *some* contextually relevant pieces of information, and especially the most significant ones.

Typically, humans have no access to the reasons on which a DSS bases its recommendations or to the roles they play in producing these recommendations. In order to be able to offer reason explanations, therefore, one would ideally be in a position to examine the actual decision-making processes of the system and to present the involved reasons and inferences accordingly. But presuming this would be naïve. More and more DSS are based on modern developments in AI. Neural networks and support vector machines, which operate on high-dimensional data spaces, seem to elude precisely this form of access and understanding of the internals, which has earned them the title of “black boxes” (Bathaee, 2018).

There are several obstacles to providing reason explanations for the recommendations of such DSS: First, there might simply be no decision process in the relevant sense. Perhaps a system learns to solve a particular task without any representation or structure at all. The concept of tacit knowledge (compare Polanyi's paradox, Autor, 2014; Polanyi, 1966) and the distinction between “knowing how” and “knowing that” (Bathaee, 2018) may be relevant in explaining how such systems can prepare recommendations and make decisions without relying on reasoning processes. Importantly, though, such systems will still offer systematic, non-random outputs relative to inputs. Otherwise they would just be random generators. But they are not—many such systems work really well, i.e., reliably provide extremely useful and fitting outputs.

Second, however, our inability to provide explanations for a DSS's outputs may be rooted in an epistemic deficiency: We simply do not gain access to hidden reasoning processes. A typical explanation of this is that the reasons and processes are represented in a distributed manner at the subsymbolic level of artificial neurons

(Goodfellow et al., 2016). But if these processes elude our access, we can certainly not easily provide them or the reasons involved therein.

And even if we could access such reasoning processes, there is a third reason why we might fail to provide the right kind of reason explanations: It is possible that the actual reasons and reasoning processes simply cannot be processed and grasped by humans, i.e., that they are incomprehensible to us (Armstrong et al., 2012). This could be the case because they are too high-dimensional to be visualized or otherwise too complicated to be suitably represented. Alternatively, such systems might use a conceptual scheme that is too different from ours to be expressible in human terms and that therefore resists translation (for doubts concerning the meaningfulness of this last claim, see Davidson, 1973).

However, these obstacles do not render the pursuit of reason explanations an impossible, hopeless endeavor. For one, the reason explanations we give for human actions are useful even though they are often approximations of far more complex processes (and may similarly face problems such as members of different cultures or linguistic communities having different conceptual schemes, or the connectionist structure of and processing in the human brain). For another, even complex reasons and reasoning processes—given they *do exist*—can in principle be approximated. A satisfactory account of how this is possible lies beyond the scope of this paper, but discussions in the philosophy of science regarding the non-factivity of understanding (Elgin, 2007), surrogate reasoning (Contessa, 2007), as well as idealization and approximation with respect to models (Frigg & Hartmann, 2020; Potochnik, 2007; Strevens, 2017) indicate a way forward. In case of their *non-existence*, we can generate sufficiently good explanations externally by methodically interpreting the systematic behavior of the DSS.<sup>35</sup>

Indeed, many existing explainability methods do something along these lines. LIME (Ribeiro et al., 2016) is a good example of this. To explain the prediction for some input, LIME approximates a complex model locally around this input by a simpler model that can then easily be explained. In other words, what is used to explain the prediction is not the original model (that may elude understanding because of its high-dimensionality), but a simpler model (with fewer dimensions) that behaves like the original model for inputs similar to the input in question. Similarly, we could generate reason explanations for a complex system by constructing a simpler system that locally approximates (relative to some observed prediction or recommendation) the original DSS. To do so, we would have to construct the simpler system in such a way that we can properly attribute reasons to its decision-making process, while staying sufficiently faithful to the behavior of the original DSS. That is to say that the simpler system has to give more or

<sup>35</sup> Remember our functional understanding of reasons and reasoning. Following Boghossian's (2014) discussion of inference, we can think of AI systems as reasoning in the following deflationary sense: As long as they are disposed, given certain premises, to come to certain conclusions (in other words, as long as they have the disposition to give certain recommendation where relevant information is available to them), they have the functional profile of a reasoner, and we will thus say that they undergo reasoning processes.

less the same recommendations for sufficiently similar inputs (for a suggestion along these lines, see Baum et al., 2017).<sup>36</sup>

Reason explanations allow humans to assess a system. For example, such explanations make it in principle possible to assess whether the system's motivating reasons are, or correspond to, normative reasons that favor the recommended action.<sup>37</sup> A well-working system responds to facts which are normative reasons.<sup>38</sup> This means that the system's recommendation will be actually favored by the facts, and the system will, in general, be robustly responsive to the facts, so it is no mere luck that it provides a good recommendation, but it does so across a broad range of situations.<sup>39</sup> By contrast, if the system is completely off track, a reason explanation of its outputs may not mention any normative reasons at all but only non-obtaining or irrelevant considerations. Even so, the explanation would be very useful to the person in the loop—for instance, by revealing that the system is malfunctioning in a particular way.

Let us apply these thoughts to our example of a hiring system.<sup>40</sup> If a hiring system recommends hiring a certain candidate for a job, a normative reason for that recommendation would be any fact that indeed obtains and that objectively favors hiring the candidate, such as the fact that she is very clever, well educated, and works accurately even under pressure. Many normative reasons may not be available to a DSS, for instance, because a CV does not fully disclose a candidate's personality and capacities. But given useful and human-processable reason explanations, a human in the loop should be able to incorporate further reasons available to her into *her* reasoning process—which is a crucial part of her role.

Now return to our disagreement case. Suppose Herbert has what he thinks are good, but not conclusive reasons for keeping April in the running, whereas the system excludes her from the top three applicants. What exactly is it that Herbert needs so he can resolve this disagreement in a non-arbitrary way? He needs to be able to figure out which party to the disagreement is in the wrong, by figuring out whether one party overlooked normative reasons that the other recognized, relied on motivating reasons that were mistaken, or gave the reasons too much or too little weight,

<sup>36</sup> Although they are not reason explanations themselves, the explanations provided by LIME can be used as good starting points to generate or extract reason explanations. Similarly to many other explainability approaches, LIME explanations highlight the most salient features for a prediction. In an image of a giraffe, for instance, we expect its long neck to be a salient feature (compared to other animals) that will be highlighted by LIME. Now, the reason for the prediction can be extracted from what is highlighted. If the reason for the prediction “giraffe” is that the image depicts an animal with a long neck, we can judge that the model works for inputs similar to this. By contrast, if the reason for the prediction “wolf” is that snow is depicted in an image, we can judge that there is an error in the model.

<sup>37</sup> For a discussion of identity and correspondence, see Mantel (2014) and Mantel (2015).

<sup>38</sup> That a system is “well-working” can mean, e.g., that it tracks reality in a reliable way, that it is accurate, or that it is robust (in a technical sense).

<sup>39</sup> See Mantel (2018) about the absence of luck in acting for normative reasons. To appropriate Fischer and Ravizza's (1998, 69) term, the system is then “reasons-receptive.”

<sup>40</sup> As may be apparent to our readers, the following toy example uses an unrealistically oversimplified DSS to emphasize the general point. In reality, DSS make their recommendations dependent on dozens, hundreds, thousands, or even millions of parameters. The art of representing adequate, but at the same time useful reason explanations in a way appropriate to the context requires not only technical but also philosophically and logically-conceptually underpinned empirical-psychological research. We say a little more about this in the outlook. Thank you to one of our reviewers for pressing us on this point.

or the like. In many situations, there are further features which modify normative reasons by disabling, attenuating, or strengthening them. They, too, need to be considered, as we will show below. In sum, the decision-maker has to be in a position to figure out whether the following possibilities are at the root of the disagreement:

*Disagreement of fact* System and decision-maker represent reality differently. They treat different propositions as facts or assign different uncertainty measures to propositions.

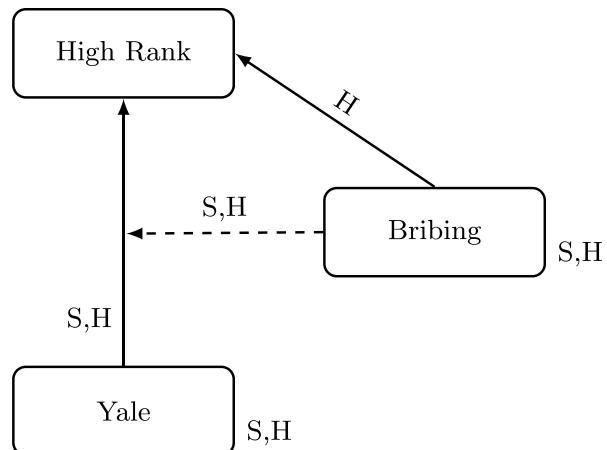
*Disagreement of relation* System and decision-maker treat different purported facts as favoring (or disfavoring) a course of action, they assign different strengths to favoring (or disfavoring) relations, or they treat purported facts as interfering with favoring (or disfavoring) relations, e.g., by disabling or attenuating them.

The human in the loop wants to check the motivating reasons on which the system relies and thus to identify disagreements of relation and disagreements of fact. For instance, there is a *disagreement of fact* if Herbert falsely believes that April has more relevant work experience than the others, whereas the system does not. If the system's rejection of April is explained by appeal to the reason that she has less work experience, this will enable him to double-check his information and to reasonably revise his original judgment (or to stick with it, if the mistake lies with the system's representation of the facts).

A *disagreement of relations* is in place, for example, if Herbert and the hiring system assume that the same facts obtain, but assign them different roles or different relations to the decision in question. This may be the case when they take the reasons in a situation to have different weights or to count in favor of different, mutually exclusive decisions; or if they disagree over whether these facts are reasons, or over whether some features modify the given reasons, as intensifiers, attenuators, enablers, or disablers. To illustrate a disagreement based on different assessments of modifiers, suppose that the DSS detects a pattern that Herbert misses: Applicants with April's (otherwise positive) traits taken together tend to move on to other, better jobs very quickly. Here, facts that would individually be great reasons to hire April, together constitute a reason *against* hiring her.

To take a more complex case (see Fig. 1), both Herbert ("H") and the DSS ("S") may be apprised of the fact that April was accepted at Yale after her mother bribed the school ("Bribing").<sup>41</sup> The system counts this fact as a disabler (dashed arrow): Given that she was accepted at Yale because of a bribe, the system does not take the fact that she studied at Yale ("Yale") to be a reason in favor of hiring her ("High Rank"). Moreover, it takes the fact that her mother bribed the school as a reason against hiring her, taking it as evidence of a lack of moral integrity. By contrast, Herbert, a person of low moral character, believes that the fact that April's mother is willing and able to use bribes to pave her daughter's way as a (prudential) reason to hire her (continuous arrow). To his mind, this fact indicates that April is from a rich, well-connected family and will therefore be an asset to the company. So, while he

<sup>41</sup> We are referencing the 2019 college admissions scam which was widely reported in the news, e.g., Jennifer Levitz and Melissa Korn (March 14, 2019), The Yale Dad Who Set Off the College-Admissions Scandal, *The Wall Street Journal*, <https://www.wsj.com/articles/the-yale-dad-who-set-off-the-college-admissions-scandal-11552588402>. Accessed April 14, 2021.

**Fig. 1** Yale bribing example

also thinks of this fact as a disabler, he treats it not as a reason against hiring her, but as a reason that favors hiring her.

Again, if Herbert receives the system's assessment that the fact that her mother bribed the school disables the fact that April went to Yale as a reason to hire her, he is in a position to integrate this knowledge into his own decision-making. For instance, he might then discount the system's recommendation because the system is blind to the importance of coming from a well-connected family; or he might come to realize that it is more important to fill this position with someone who made it into an excellent university without bribery, and comply with the system's recommendation. Either way, he will meet the epistemic condition with respect to his decision, and bear moral responsibility for it.

Generally speaking, disagreements are typically resolved by taking into account the reasons of the other party. The decision-maker needs a grasp of what reasons the system operated with and how it treated pieces of information, e.g., as reasons or as disablers. This is to say, a reason explanation needs to state explicitly what pieces of information served as reasons for or against a certain recommendation and what pieces of information served as modifiers of reasons. Furthermore, the explanation needs to include the strengths of these reasons. If the decision-maker has access to this information, he can reassess his information about the facts as well as his treatment of them as playing different roles such as those of reasons, disablers, attenuators, etc. He can then come to an all-things-considered decision that integrates all relevant facts in a coherent way, weighing the relevant reasons against each other, and he is then in a position to know his decision under the relevant descriptions and, thus, to be morally responsible for the decision.<sup>42</sup>

Reason explanations are the explanations we typically use to communicate reasons. By contrast, other forms of explanations would seem to make resolving disagreements much harder. Imagine that Herbert is provided with the following explanation of why

<sup>42</sup> It should be noted that—as the reader may already suspect—Herbert's use of the system's reason explanations not only makes him responsible in the sense of accountable but also puts him in a position to be answerable (see Shoemaker 2011, 616) for the resulting decision. For it enables him to cite the reasons that he took to justify the decision when he made it. Our focus here, however, is on the role that explainability plays for responsibility not in the sense of answerability, but of accountability.

the system recommended against April: If her mother had not bribed the school, it would have recommended to hire April. Such a *counterfactual* explanation, as suggested by Wachter et al. (2018), indicates that the facts mentioned in the explanation were taken either as reasons against hiring April, or as disablers of other reasons to hire her (and in the example, both are the case). But as this example illustrates, he may still be unable to tell which of the two roles a fact played (reason against or disabler, or both)—with one role he agrees, with the other he does not—and it will take extra work for him to assign the facts their proper roles and to integrate them correctly into his own reasoning.

## 6 Open Questions and Future Work

In this paper, we have argued that, to close responsibility gaps, we often need a human in the loop who is in a position to bear direct responsibility for her—AI-supported—decisions. However, for a human in the loop to be in a position of directly responsible decision-making, she needs to have the right kind of epistemic access to relevant features of her action. We have argued that the epistemic condition on moral responsibility often cannot be met by the human in the loop if she has no access to the system’s motivating reasons for its recommendation. We have explained how meeting the epistemic condition translates to certain abilities in practice, first and foremost to the ability to recognize and resolve disagreements of different kinds between man and machine. And we have argued that reason explanations are theoretically well-suited to restore epistemic access, supplying a background picture of motivating and normative reasons from the philosophy of action, which we started to transfer to decision support systems and their recommendations.

However, all this can only be a starting point. Several empirical and technical tasks remain on the path to useful machine-generated reason explanations. In a further empirical step, one could ask *how many* and *which* motivating reasons need to be provided (especially when a DSS processes a great amount of information) and *how they need to be presented* in the explanation of a recommendation such that the human can best use the explanation. This includes the issue of how the strength of reasons or different roles such as disabling, attenuating, or intensifying should be represented. In explaining a system’s recommendations, what is needed is an explanation that users can understand and often one that they can comprehend quickly. When humans give explanations, they intuitively present information selectively and focus on the information that seems relevant in the context of a given question. Providing more information than necessary can be distracting, and it leaves the recipient of the explanation the time-consuming task of singling out the bits that are most relevant. This can be counterproductive.<sup>43</sup> Hence, the explanation presented initially would ideally involve only the most relevant motivating considerations, while flagging their respective roles. Less relevant motivating reasons would be provided only upon a request to give more detailed information. But which reasons are relevant and why? This calls for further, especially psychological and normative research.

<sup>43</sup> For instance, more information about a digital application process can easily undermine the organizational attractiveness (Langer et al., 2018). It seems not too far-fetched to assume similar counterproductive effects due to too many or poorly presented reasons.

Relatedly, it seems problematic to try to provide *general* principles of which reasons will be the most relevant elements of a reason explanation. For the relevance of a reason is not determined solely by its significance within the specific reasoning process, but might well be a function also of the aims or background knowledge of the human who receives the explanation. This seems to call for an interactive way of explaining that allows the human to dive deeper into the why, a typical problem for human–computer interaction.<sup>44</sup>

This brings us to the question of how, quite generally, reason explanations of AI systems' outputs should be represented to their addressees. What is a suitable data structure for reason explanations? We not only need a way to represent reasons, but also their relations and quantitative information like their weight or potentially involved uncertainty. Formal, graph-based approaches to reasons (Horty, 2012) as well as argumentation and dialectical frameworks (Amgoud & Prade, 2009; Baum et al., 2018, 2019; Dung, 1995) might lead the way. Further research along these lines calls for input from both theoretical and practical computer scientists.

In short, we believe that the endeavor of equipping decision support systems with the ability of giving reason explanations is not only imperative, but opens up several interesting and highly interdisciplinary lines of research for the future.

**Acknowledgements** We thank audiences at the Rhine-Ruhr Epistemology Meeting; at the lecture series Gratwanderung Künstliche Intelligenz (Dortmund); at the Colloquium of the Department of Values, Technology and Innovation (Delft); at the reading group for theoretical philosophy (Dortmund); at the EIS Kolloquium (Saarbrücken/Dortmund); at the research colloquium of the DCLPS (Düsseldorf); and at the reading group for philosophy and technology (Frankfurt School of Finance & Management) for helpful discussion. Special thanks go to Benjamin Kiesewetter, Sebastian Köhler, Sara Mann, Leonhard Menges, Filippo Santoni de Sio, Sarah Sterz, Christine Tiefensee, Johannes Weyer, and Simon Wimmer, for lots of great input. Finally, we are grateful for the in-depth comments of two anonymous reviewers, which helped us to greatly improve our argument.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work was partially supported by VolkswagenStiftung as part of Grant AZ 98510, 98512, and 98514 EIS—Explainable Intelligent Systems (see <https://explainable-intelligent.systems/>), and partially funded by DFG grant 389792660 as part of TRR 248, CPEC (see <https://perspicuous-computing.science>).

## Declarations

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

<sup>44</sup> Langer et al. (2021b) and Schlicker et al. (2021) provide some insights into context factors that may impact what is required of an explanation to be of use to a human in a certain context.

## References

- Alvarez, M. (2010). Kinds of Reasons: An Essay in the Philosophy of Action. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199550005.001.0001>
- Alvarez, M. (2017). Reasons for Action: Justification, Motivation, Explanation. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/>
- Amgoud, L., & Prade, H. (2009). Using Arguments for Making and Explaining Decisions. *Artificial Intelligence*, 173(3–4), 413–436. <https://doi.org/10.1016/j.artint.2008.11.006>
- Anscombe, G. E. M. (1962). *Intention*. Blackwell Press.
- Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4), 299–324. <https://doi.org/10.1007/s11023-012-9282-2>
- Asaro, P. M. (2015). The Liability Problem for Autonomous Artificial Agents. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposia*, 190–194. AAAI.
- Autor, D. (2014). Polanyi's Paradox and the Shape of Employment Growth. *National Bureau of Economic Research Working Paper Series*. <https://doi.org/10.3386/w20485>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7). <https://doi.org/10.1371/journal.pone.0130140>
- Bathaei, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*, 31(2), 889–938. <https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaei.pdf>
- Baum, K., Köhl, M. A., & Schmidt, E. (2017). Two Challenges for CI Trustworthiness and How to Address Them. In *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*. <https://doi.org/10.18653/v1/w17-3701>
- Baum, K., Hermanns, H., & Speith, T. (2018). From Machine Ethics to Machine Explainability and Back. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*. <https://www.powver.org/publications/TechRepRep/ERC-POWVER-TechRep-2018-02.pdf>
- Baum, K., Hermanns, H., & Speith, T. (2019). Towards a Framework Combining Machine Ethics and Machine Explainability. In *Proceedings of the 3rd Workshop on formal reasoning about Causation, Responsibility, and Explanations in Science and Technology (CREST)* (pp. 34–49). <https://doi.org/10.4204/epics.286.4>
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). ‘It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3173951>
- Boghossian, P. (2014). What Is Inference? *Philosophical Studies*, 169(1), 1–18. <https://doi.org/10.1007/s11098-012-9903-x>
- Cave, S., Nyrop, R., Vold, K., & Weller, A. (2018). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107(3), 562–574. <https://doi.org/10.1109/jproc.2018.2865996>
- Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *IEEE 29th International Requirements Engineering Conference (RE)* (pp. 197–208). IEEE. <https://doi.org/10.1109/RE51729.2021.00025>
- Chockler, H., & Halpern, J. Y. (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22, 93–115. <https://doi.org/10.1613/jair.1391>
- Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Coectelbergh, M. (2020). AI Ethics. *The MIT Press*. <https://doi.org/10.7551/mitpress/12549.001.0001>
- Contessa, G. (2007). Scientific Representation, Interpretation, and Surrogate Reasoning. *Philosophy of Science*, 74(1), 48–68. <https://doi.org/10.1086/519478>
- Cooper, B. (2016). Intersectionality. In Disch, L., & Hawkesworth, M. (Eds.), *The Oxford Handbook of Feminist Theory* (pp. 385–406). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199328581.013.20>
- Dancy, J. (2000). Practical Reality. *Oxford University Press*. <https://doi.org/10.1093/0199253056.001.0001>

- Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, 60(23), 685–700. <https://doi.org/10.2307/2023177>
- Davidson, D. (1973). Radical Interpretation. *Dialectica*, 27(3–4), 313–328. <https://doi.org/10.1111/j.1746-8361.1973.tb00623.x>
- de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, 31(4), 525–541. <https://doi.org/10.1007/s13347-017-0293-z>
- Duff, R. A. (2007). *Answering for Crime*. Hart Publishing.
- Duff, R. A. (2019). Moral and Criminal Responsibility: Answering and Refusing to Answer. In Coates, D. J., & Tognazzini N. A. (Eds.), *Oxford Studies in Agency and Responsibility Volume 5* (pp. 165–190). Oxford University Press. <https://doi.org/10.1093/oso/9780198830238.003.0009>
- Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games. *Artificial Intelligence*, 77(2), 321–357. [https://doi.org/10.1016/0004-3702\(94\)00041-x](https://doi.org/10.1016/0004-3702(94)00041-x)
- Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a “Right to Explanation” Is Probably Not the Remedy You Are Looking for. *Duke Law & Technology Review*, 16, 18–84. <https://doi.org/10.2139/ssrn.2972855>
- Elgin, C. Z. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42. <https://doi.org/10.1007/s11098-006-9054-z>
- Fehr, E., & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature*, 415(6868), 137–140. <https://doi.org/10.1038/415137a>
- Fischer, J. M., & Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility. *Cambridge University Press*. <https://doi.org/10.1017/CBO9780511814594>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Frigg, R., & Hartmann, S. (2020). Models in Science. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>
- Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, 33(4), 111–117. <https://doi.org/10.1215/07402775-3813015>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <https://www.deeplearningbook.org>
- Halpern, J. Y., & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4), 843–887. <https://doi.org/10.1093/bjps/axi147>
- Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sciences*, 54(2), 269–287. <https://doi.org/10.1007/s11077-020-09414-y>
- Hieronymi, P. (2011). XIV-Reasons for Action. *Proceedings of the Aristotelian Society*, 111(3), 407–427. <https://doi.org/10.1111/j.1467-9264.2011.00316.x>
- Horty, J. F. (2012). Reasons as Defaults. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199744077.001.0001>
- Johnson, D. G. (2015). Technology with No Human Responsibility? *Journal of Business Ethics*, 127(4), 707–715. <https://doi.org/10.1007/s10551-014-2180-1>
- Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169–175). IEEE. <https://doi.org/10.1109/REW53955.2021.00031>
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677). <https://proceedings.mlr.press/v80/kim18d/kim18d.pdf>
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. In *IEEE 27th International Requirements Engineering Conference (RE)* (pp. 363–368). IEEE <https://doi.org/10.1109/RE.2019.00046>

- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. & Yu, Harlan (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165, <https://ssrn.com/abstract=2765268>.
- Langer, M., König, C. J., & Fitili, A. (2018). Information as a Double-Edged Sword: The Role of Computer Experience and Information on Applicant Reactions towards Novel Technologies for Personnel Selection. *Computers in Human Behavior*, 81, 19–30. <https://doi.org/10.1016/j.chb.2017.11.036>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021a). What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence*, 296,. <https://doi.org/10.1016/j.artint.2021.103473>
- Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021b). Spare Me the Details: How the Type of Information about Automated Interviews Influences Applicant Reactions. *International Journal of Selection and Assessment*, 29(2), 154–169. <https://doi.org/10.1111/ijsa.12325>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J. Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., & Thrun, S. (2011). Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)* (pp. 163–168). IEEE. <https://doi.org/10.1109/IVS.2011.5940562>
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Mantel, S. (2014). No Reason for Identity: On the Relation between Motivating and Normative Reasons. *Philosophical Explorations*, 17(1), 49–62. <https://doi.org/10.1080/13869795.2013.815261>
- Mantel, S. (2015). Worldly Reasons: An Ontological Inquiry into Motivating Considerations and Normative Reasons. *Pacific Philosophical Quarterly*, 98(S1), 5–28. <https://doi.org/10.1111/papq.12094>
- Mantel, S. (2018). Determined by Reasons. *Routledge*. <https://doi.org/10.4324/9781351186353>
- Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- McKenna, M. (2012). Conversation and Responsibility. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199740031.001.0001>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafiyan, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, 22(2), 103–115. <https://doi.org/10.1007/s10676-019-09519-w>
- Mele, A. R. (2021). Direct Versus Indirect: Control, Moral Responsibility, and Free Action. *Philosophy and Phenomenological Research*, 102(3), 559–573. <https://doi.org/10.1111/phpr.12680>
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)* (pp. 36–42).
- Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), <https://doi.org/10.1177/2053951716679679>
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65, 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>
- Nissenbaum, H. (1996). Accountability in a Computerized Society. *Science and Engineering Ethics*, 2(1), 25–42. <https://doi.org/10.1007/bf02639315>

- Noorman, M. (2020). Computing and Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>
- Pereboom, D. (2014). Free Will, Agency, and Meaning in Life. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199685516.001.0001>
- Perel, M., & Elkin-Koren, N. (2016). Accountability in Algorithmic Copyright Enforcement. *Stanford Technology Law Review*, 19, 473–533. <https://doi.org/10.2139/ssrn.2607910>
- Polanyi, M. (1966). *The Tacit Dimension*. Routledge and Kegan Paul.
- Potochnik, A. (2007). Optimality Modeling and Explanatory Generality. *Philosophy of Science*, 74(5), 680–691. <https://doi.org/10.1086/525613>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
- Rosen, G. (2003). IV-Culpability and Ignorance. *Proceedings of the Aristotelian Society*, 103(1), 61–84. <https://doi.org/10.1111/j.0066-7372.2003.00064.x>
- Rudy-Hiller, F. (2018). The Epistemic Condition for Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/>
- Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5. <https://doi.org/10.3389/frobt.2018.00015>
- Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 1–28. <https://doi.org/10.1007/s13347-021-00450-x>
- Scanlon, T. M. (2008). Moral Dimensions. *Harvard University Press*. <https://doi.org/10.4159/978074043145>
- Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to Expect from Opening up ‘Black Boxes’? Comparing Perceptions of Justice Between Human and Automated Agents. *Computers in Human Behavior*, 122. <https://doi.org/10.1016/j.chb.2021.106837>
- Schmidt, E. (2018). Normative Reasons for Mentalism. In Kyriacou, C., & McKenna, R. (Eds.), *Metaepistemology: Realism and Anti-Realism* (pp. 97–120). Palgrave Macmillan. [https://doi.org/10.1007/978-3-319-93369-6\\_5](https://doi.org/10.1007/978-3-319-93369-6_5)
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). <https://doi.org/10.1109/iccv.2017.74>
- Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121(3), 602–632. <https://doi.org/10.1086/659003>
- Shoemaker, D. (2012). Blame and Punishment. In Coates, D. J., & Tognazzini, N. A. (Eds.), *Blame: Its Nature and Norms* (pp. 100–118). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199860821.003.0006>
- Shoemaker, D. (2013). On Criminal and Moral Responsibility. *Oxford Studies in Normative Ethics*, 3, 154–178. <https://doi.org/10.1093/acprof:oso/9780199685905.003.0008>
- Shoemaker, D. (2015). Responsibility from the Margins. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780198715672.001.0001>
- Smith, A. M. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*, 115(2), 236–271. <https://doi.org/10.1086/426957>
- Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, 24(1), 62–77. <https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Sterz, S., Baum, K., Lauber-Rönsberg, A., & Hermanns, H. (2021). Towards Perspicuity Requirements. In *IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 159–163). IEEE. <https://doi.org/10.1109/REW53955.2021.00029>
- Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48, 1–25.
- Strevens, M. (2017). How Idealizations Provide Understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Essays in Epistemology and the Philosophy of Science* (pp. 37–49). Routledge.
- Talbert, M. (2019). Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/>
- Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, 74(4), 905–916. <https://doi.org/10.2307/1954312>

- van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). Moral Responsibility and the Problem of Many Hands. *Routledge*. <https://doi.org/10.4324/9781315734217>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ixp005>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 842–861. <https://doi.org/10.2139/ssrn.3063289>
- Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.
- Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, 24(2), 227–248. <https://doi.org/10.5840/philtopics199624222>
- Zarsky T. (2013). Transparency in Data Mining: From Theory to Practice. In Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.), *Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases* (pp. 301–324). Springer. [https://doi.org/10.1007/978-3-642-30487-3\\_17](https://doi.org/10.1007/978-3-642-30487-3_17)
- Zimmerman, M. J. (1997). Moral Responsibility and Ignorance. *Ethics*, 107(3), 410–426. <https://doi.org/10.1086/233742>
- Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, 32(4), 661–683. <https://doi.org/10.1007/s13347-018-0330-6>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Software doping analysis for human oversight

Sebastian Biewer<sup>1</sup> · Kevin Baum<sup>1,2,3</sup> · Sarah Sterz<sup>1</sup> · Holger Hermanns<sup>1</sup> ·  
Sven Hetmank<sup>4</sup> · Markus Langer<sup>5</sup> · Anne Lauber-Rönsberg<sup>4</sup> · Franz Lehr<sup>4</sup>

Received: 22 December 2022 / Accepted: 11 January 2024  
© Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

This article introduces a framework that is meant to assist in mitigating societal risks that software can pose. Concretely, this encompasses facets of software doping as well as unfairness and discrimination in high-risk decision-making systems. The term *software doping* refers to software that contains surreptitiously added functionality that is against the interest of the user. A prominent example of software doping are the tampered emission cleaning systems that were found in millions of cars around the world when the diesel emissions scandal surfaced. The first part of this article combines the formal foundations of software doping analysis with established probabilistic falsification techniques to arrive at a black-box analysis technique for identifying undesired effects of software. We apply this technique to emission cleaning systems in diesel cars but also to high-risk systems that evaluate humans in a possibly unfair or discriminating way. We demonstrate how our approach can assist humans-in-the-loop to make better informed and more responsible decisions. This is to promote effective human oversight, which will be a central requirement enforced by the European Union's upcoming AI Act. We complement our technical contribution with a juridically, philosophically, and psychologically informed perspective on the potential problems caused by such systems.

**Keywords** Software doping · Artificial intelligence · Algorithmic fairness · Probabilistic falsification · Adequate trust · Human oversight

## 1 Introduction

Software is the main driver of innovation of our times. Software-defined systems are permeating our communication, perception, and storage technology as well as our personal interactions with technical systems at an unprecedented pace. “*Software-defined everything*” is among the hottest buzzwords in IT today [78, 121].

- 
- ✉ Sebastian Biewer  
biewer@depend.uni-saarland.de
  - ✉ Kevin Baum  
kevin.baum@dfki.de
  - ✉ Sarah Sterz  
sterz@depend.uni-saarland.de

Extended author information available on the last page of the article

At the same time, we are doomed to trust these systems, despite being unable to inspect or look inside the software we are facing: The owners of the physical hull of ‘everything’ are typically not the ones owning the software defining ‘everything’, nor will they have the right to look at what and how ‘everything’ is defined. This is because commercial software typically is protected by intellectual property rights of the software manufacturer. This prohibits any attempt to disassemble the software or to reconstruct its inner working, albeit it is the very software that is forecasted to be defining ‘everything’. The use of machine-learnt software components amplifies the problem considerably by adding opacity of its own kind. Since commercial interests of the software manufacturers seldomly are aligned with the interest of end users, the promise of ‘software-defined everything’ might well become a dystopia from the perspective of individual digital sovereignty. In this article, we address two of the most pressing incarnations of problematic software behaviour.

### *Diesel emissions scandal*

A massive example of software-defined collective damage is the diesel emissions scandal. Over a period of more than 10 years, millions of diesel-powered cars have been equipped with illegal software that altogether polluted the environment for the sake of commercial advantages of the car manufacturers. At its core, this was made possible by the fact that only a single, precisely defined test setup was put in place for checking conformance with exhaust emissions regulations. This made it a trivial software engineering task to identify the test particularities and to turn off emission cleaning outside these particular conditions. This is an archetypal instance of software doping.

Software doping can be formally characterised as a violation of a *cleanness* property of a program [10, 32]. A detailed and comparative account of meaningful cleanliness definitions related to software doping is available [16, Chapter 3]. One cleanliness notion that has proven suitable to detect diesel emissions doping is *robust cleanliness* [16, 19]. It is based on the assumption that there is some well-defined and agreed standard input/output behaviour of the system which the definition extends to the vicinity around the inputs and outputs close to the standard behaviour. The precise specification of “vicinity” and of “standard behaviour” is assumed to be part of a *contract* between software manufacturer and user. That contract entails the standard behaviour, distance functions for input and output values, and distance thresholds to define the input and output vicinity, respectively. With this, a system behaviour is considered clean, if its output is (or stays) in the output vicinity of the standard, unless the input is (or moves) outside the standard’s input vicinity (see Fig. 1).

**Example 1** Every car model that is to enter the market in the European Union (and other countries) must be compliant with local regulations. As part of this homologation process, common to all of these regulations is the need for executing a test under precisely defined lab conditions, carried out on a chassis dynamometer. In this, the car has to follow a speed profile, which is called *test cycle* in regulations. At the time when the diesel scandal surfaced, the *New European Driving Cycle* (NEDC) [128] was the single test cycle used in the European Union. It has by now been replaced by the *Worldwide harmonized Light vehicles Test Cycle* (WLTC) [124] in many countries. We refer to previous work for more details [16, 19, 22]. From a perspective of fraud prevention, having only a single test cycle is a major weakness of the homologation procedure. Robust cleanliness can overcome this problem. It admits the consideration of driving profiles that stay in the bounded vicinity of one of several standardised test cycle (i.e., NEDC as well as WLTC), while enforcing bounds on the deviations regarding exhaust emission.

### *Discrimination mitigation*

71 Another set of exemplary scenarios we consider in this article are *high-risk* AI systems, sys-  
72 tems empowered by AI technology whose functioning may introduce risks to health, safety,  
73 or fundamental rights of human individuals. The European Union is currently developing the  
74 *AI Act* [40, 41] that sets out to mitigate many of the risks that such systems pose. Applica-  
75 tion areas of concern include credit approval ([95]), decisions on visa applications ([84]),  
76 admissions to higher education ([27, 133]), screening of individuals in predictive policing  
77 ([58]), selection in HR ([92–94]), judicial decisions (as with COMPAS [3, 30, 34, 72]),  
78 tenant screening ([115]), and more. In many of these areas, there are legitimate interests and  
79 valid reasons for using well-understood AI technology, although the risks associated with  
80 their use to date is manifold.

81 It is widely recognised that discrimination by unfair classification and regression models is  
82 one particularly important risk. As a result, a colourful zoo of different operationalisations of  
83 unfairness has emerged [96, 131], which should be seen less as a set of competing approaches  
84 and more as mutually complementary [52]. At the same time, a consensus is emerging that  
85 human oversight is an important piece of the puzzle for mitigating and minimising societal  
86 risks of AI [59, 83, 129]. Accordingly, that principle made it into recent drafts of legislation  
87 including the European AI Act [40, 41] or certain US state laws [132].

88 The generic approach we develop for software-doping analysis turns out to be powerful  
89 enough to provide automated assistance for human overseers of high-risk AI systems. Apart  
90 from spelling out the necessary refocusing, we illustrate the challenge that our work helps  
91 to overcome by an exemplary, albeit hypothetical admission system for higher education  
92 (inspired by [27, 133]).

93 **Example 2** A large university assigns scores to applicants aiming to enter their computer  
94 science PhD program. The scores are computed using an automated, model-based procedure  
95 P which is based on three data points: the position of the applicant's last graduate institution  
96 in an official, subject-specific ranking, the applicant's most recent grade point average (GPA),  
97 and their score in a subject-specific standardised test taken as part of the application procedure.  
98 The system then automatically computes a score for the candidate based on an estimation of  
99 how successful it expects them to be as students. A dedicated university employee, Unica is in  
100 charge of overseeing the individual outcomes of P and is supposed to detect cases where the  
101 output of P is or appears flawed. The university pays especial attention to fairness in the scoring  
102 procedure, so Unica has to watch out to any signs of potential unfairness. Unica is supposed  
103 to desk-reject candidates whose scores are below a certain, predefined threshold—unless she  
104 finds problems with P's scoring. Without any additional support, Unica, as human overseer  
105 in the loop, must manually check all cases for signs of unfairness as they are processed. This  
106 can be a tedious, complicated, and error-prone task and as such constitutes an impediment  
107 for the assumed scalability of the automated scoring process to high numbers of applicants.  
108 Therefore, she at least requires tool support that assists her in detecting when something is  
109 off about the scoring of individual applicants.

110 This support can be made real by exploiting the technical contributions of this article, in  
111 terms of a runtime monitor that provides automated assistance to the human oversight and  
112 itself is based on the probabilistic falsification technique we develop. As we will explain,  
113 func-cleanliness, a variant of cleanliness, is a suitable basis for rolling out runtime monitors for  
114 such high-risk systems, that are able to detect and flag discrimination or unfair treatment of  
115 humans.

116 The contributions made by this article are threefold.

117 *Detecting software doping using probabilistic falsification.* The paper starts off by develop-  
118 ing the theory of robust cleanliness and func-cleanliness. We provide characterisations in the

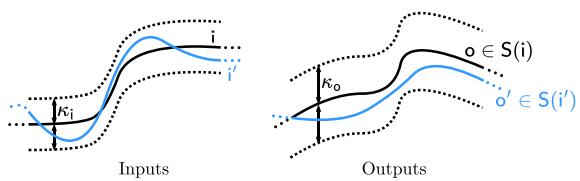
119 temporal logics HyperSTL and STL, that are then used for an adaptation of existing prob-  
120 abilistic falsification techniques [1, 49]. Altogether, this reduces the problem of software  
121 doping detection to the problem of falsifying the logical characterisation of the respective  
122 cleanliness definition.

123 *Falsification-based test input generation.* Recent work [19] proposes a formal framework  
124 for robust cleanliness testing, with the ambition of making it usable in practice, namely for  
125 emissions tests conducted with a real diesel car on a chassis dynamometer. However, that  
126 approach leaves open how to perform test input selection in a meaningful manner. The prob-  
127 abilistic falsification technique presented in this article attacks this shortcoming. It supports  
128 the testing procedure by guiding it towards test inputs that make the robust cleanliness tests  
129 likely to fail.

130 *Promoting effective human oversight.* We discuss and demonstrate how the technical contribu-  
131 tions of this paper contribute to effective human oversight of high-risk systems, as required by  
132 the current proposal of the AI act. The hypothetical university admission scenario introduced  
133 above will serve as a demonstrator for shedding light on the applicability of our approach  
134 as well as the the principles behind it. On a technical level, we provide a runtime monitor  
135 for individual fairness based on probabilistic falsification of func-cleanliness. On a conceptual  
136 level, we consider it important to clarify which duties come with the usage of such a system;  
137 from a *legal* perspective, particularly considering the AI Act, substantiated by considering the  
138 *ethical* dimension from a philosophical perspective, and from a *psychological* perspective,  
139 particularly deliberating on how the overseeing can become *effective*.

140 This paper is based on a conference publication [17]. Relative to that paper, the develop-  
141 ment of the theory here is more complete and now includes temporal logic characterisations  
142 for func-cleanliness. On the conceptual side, this article adds a principled analysis of the appli-  
143 cability of func-cleanliness to effective human oversight, spelled out in the setting of admission  
144 to higher education. We live up to the societal complexity of this new example and provide  
145 an interdisciplinary situation analysis and an interdisciplinary assessment of our proposed  
146 solution. Accordingly, although the technical realisation is based on the probabilistic falsifi-  
147 cation approach outlined in this article, our solution is substantially more thoughtful than a  
148 naive instantiation of the falsification framework.

149 This article is structured as follows. Section 2 provides the preliminaries for the contri-  
150 butions in this article. Section 3 develops the theoretical foundations necessary to use the  
151 concept of probabilistic falsification with robust cleanliness and func-cleanliness. Section 4  
152 demonstrates how the probabilistic falsification approach can be combined with the previ-  
153 ously proposed testing approach [19] for robust cleanliness, with a focus on tampered emission  
154 cleaning systems of diesel cars. Section 5 develops the technical realisation of a fairness mon-  
155 itor based on func-cleanliness for high-risk systems. Section 6 evaluates the fairness monitor  
156 from the perspective of the disciplines philosophy, psychology, and law. Finally, Sect. 7  
157 summarises the contributions of this article and discusses limitations of our approaches. The  
158 appendix of this article contains additional technical details, proofs, and further philosophical  
159 and juridical explanations.

**Fig. 1** Robust cleanliness intuition

## 2 Background

### 2.1 Software doping

After early informal characterisations of *software doping* [10, 13], D’Argenio et al. [32] propose a collection of formal definitions that specify when a software is *clean*. The authors call a software *doped* (w.r.t. a cleanliness definition) whenever it does not satisfy such cleanliness definition. We focus on *robust cleanliness* and *func-cleanliness* in this article [32].

We define by  $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \mid x \geq 0\}$  the set of non-negative real numbers, by  $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$  the set of *extended reals* [104], and by  $\bar{\mathbb{R}}_{\geq 0} := \mathbb{R}_{\geq 0} \cup \{\infty\}$  the set of the non-negative extended real numbers. We say that a function  $d : X \times X \rightarrow \bar{\mathbb{R}}_{\geq 0}$  is a *distance function* if and only if it satisfies  $d(x, x) = 0$  and  $d(x, y) = d(y, x)$  for all  $x, y \in X$ . We let  $\sigma[k]$  denote the  $k$ th literal of the finite or infinite word  $\sigma$ .

#### Reactive Execution Model

We can view a nondeterministic reactive program as a function  $S : \text{In}^\omega \rightarrow 2^{(\text{Out}^\omega)}$  perpetually mapping inputs  $\text{In}$  to sets of outputs  $\text{Out}$  [32]. To formally model contracts that specify the concrete configuration of robust cleanliness or func-cleanliness, we denote by  $\text{StdIn} \subseteq \text{In}^\omega$  the input space of the system designated to define the standard behaviour, and by  $d_{\text{In}} : (\text{In} \times \text{In}) \rightarrow \bar{\mathbb{R}}_{\geq 0}$  and  $d_{\text{Out}} : (\text{Out} \times \text{Out}) \rightarrow \bar{\mathbb{R}}_{\geq 0}$  distance functions on inputs, respectively outputs.

For robust cleanliness, we additionally consider two constants  $\kappa_i, \kappa_o \in \bar{\mathbb{R}}_{\geq 0}$ .  $\kappa_i$  defines the maximum distance that a non-standard input must have to a standard input to be considered in the cleanliness evaluation. For this evaluation,  $\kappa_o$  defines the maximum distance between two outputs such that they are still considered sufficiently close. Intuitively, the contract defines tubes around standard inputs and their outputs. For example, in Fig. 1,  $i$  is a standard input and  $d_{\text{In}}$  and  $\kappa_i$  implicitly define a  $2\kappa_i$  wide tube around  $i$ . Every input  $i'$  that is within this tube will be evaluated on its outputs. Similarly,  $d_{\text{Out}}$  and  $\kappa_o$  define a tube around each of the outputs of  $i$ . An output for  $i'$  that is within this tube satisfies the robust cleanliness condition. Together, the above objects constitute a formal contract  $\mathcal{C} = \langle \text{StdIn}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$ . Robust cleanliness is composed of two separate definitions called l-robust cleanliness and u-robust cleanliness. Assuming a fixed standard behaviour of a system, l-robust cleanliness imposes a lower bound on the non-standard outputs that a system must exhibit, while u-robust cleanliness imposes an upper bound. Such lower and upper bound considerations are necessary because of the potential nondeterministic behaviour of the system; for deterministic systems the two notions coincide. We remark that in this article we are using past-forgetful distance functions and the trace integral variants of robust cleanliness and func-cleanliness (see Biewer [16, Chapter 3] for details).

**Definition 1** A nondeterministic reactive program  $S : \text{In}^\omega \rightarrow 2^{(\text{Out}^\omega)}$  is *robustly clean* w.r.t. contract  $\mathcal{C} = \langle \text{StdIn}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  if for every standard input  $i \in \text{StdIn}$  and input sequence  $i' \in \text{In}^\omega$  it is the case that

- 197 1. for every  $\mathbf{o} \in S(i)$ , there exists  $\mathbf{o}' \in S(i')$ , such that for every index  $k \in \mathbb{N}$ , if  
 198      $d_{In}(i[j], i'[j]) \leq \kappa_i$  for all  $j \leq k$ , then it holds that  $d_{Out}(\mathbf{o}[k], \mathbf{o}'[k]) \leq \kappa_o$ ,  
     *(l-robust cleanliness)*  
 199
- 200 2. for every  $\mathbf{o}' \in S(i')$ , there exists  $\mathbf{o} \in S(i)$ , such that for every index  $k \in \mathbb{N}$ , if  
 201      $d_{In}(i[j], i'[j]) \leq \kappa_i$  for all  $j \leq k$ , then it holds that  $d_{Out}(\mathbf{o}[k], \mathbf{o}'[k]) \leq \kappa_o$ .  
     *(u-robust cleanliness)*  
 202

203 We will in the following refer to Definition 1.1 for l-robust cleanliness and Definition 1.2 for  
 204 u-robust cleanliness. Intuitively, l-robust cleanliness enforces that whenever an input  $i'$  remains  
 205 within  $\kappa_i$  vicinity around the standard input  $i$ , then for every standard output  $\mathbf{o} \in S(i)$ , there  
 206 must be a non-standard output  $\mathbf{o}' \in S(i')$  that is in  $\kappa_o$  proximity of  $\mathbf{o}$ . Referring to Fig. 1,  
 207 every  $i'$  in the tube around  $i$  must produce for every standard output  $\mathbf{o} \in S(i)$  at least one  
 208 output  $\mathbf{o}' \in S(i')$  that resides in the  $\kappa_o$ -tube around  $\mathbf{o}$ . In other words, for non-standard inputs  
 209 the system must not lose behaviour that it can exhibit for a standard input in  $\kappa_i$  proximity.

210 For u-robust cleanliness the standard and non-standard output switch roles. It enforces that  
 211 whenever an input  $i'$  remains within  $\kappa_i$  vicinity around the standard input  $i$ , then for every  
 212 output  $\mathbf{o}' \in S(i')$  the system can exhibit for this non-standard input, there must be a standard  
 213 output  $\mathbf{o} \in S(i)$  that is in  $\kappa_o$  proximity of  $\mathbf{o}'$ . Referring to Fig. 1, every  $i'$  in the tube around  $i$   
 214 must only produce outputs  $\mathbf{o}' \in S(i')$  that are in the  $\kappa_o$ -tube of at least one  $\mathbf{o} \in S(i)$ . In other  
 215 words, for non-standard inputs within  $\kappa_i$  proximity of a standard input the system must not  
 216 introduce new behaviour, i.e., it must not exhibit an output that is further than  $\kappa_o$  away from  
 217 the set of standard outputs.

218 A generalisation of robust cleanliness is func-cleanliness. A cleanliness contract for func-  
 219 cleanliness replaces the constants  $\kappa_i$  and  $\kappa_o$  by a function  $f : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  inducing a dynamic  
 220 threshold for output distances based on the distance between the inputs producing such  
 221 outputs.

222 **Definition 2** A nondeterministic reactive system  $S$  is *func-clean* w.r.t. contract  $C =$   
 223  $\langle StdIn, d_{In}, d_{Out}, f \rangle$  if for every standard input  $i \in StdIn$  and input sequence  $i' \in In^\omega$  it  
 224 is the case that

- 225 1. for every  $\mathbf{o} \in S(i)$ , there exists  $\mathbf{o}' \in S(i')$ , such that for every index  $k \in \mathbb{N}$ ,  
 226      $d_{Out}(\mathbf{o}[k], \mathbf{o}'[k]) \leq f(d_{In}(i[k], i'[k]))$ ,  
     *(l-func-cleanness)*  
 227 2. for every  $\mathbf{o}' \in S(i')$ , there exists  $\mathbf{o} \in S(i)$ , such that for every index  $k \in \mathbb{N}$ ,  
 228      $d_{Out}(\mathbf{o}[k], \mathbf{o}'[k]) \leq f(d_{In}(i[k], i'[k]))$ .  
     *(u-func-cleanness)*

229 We will in the following refer to Definition 2.1 for l-func-cleanliness and Definition 2.2 for  
 230 u-func-cleanliness.

231 For the fairness monitor in Sect. 5 we will use a simpler variant of func-cleanliness for  
 232 deterministic sequential programs. Since  $P$  is deterministic, the lower and upper bound  
 233 requirements coincide, yielding the following simplified definition.

234 **Definition 3** A deterministic sequential program  $P$  is *func-clean* w.r.t. contract  $C =$   
 235  $\langle StdIn, d_{In}, d_{Out}, f \rangle$  if for every standard input  $i \in StdIn$  and input  $i' \in In$ , it holds that  
 236  $d_{Out}(P(i), P(i')) \leq f(d_{In}(i, i'))$ .

### 237 **Mixed-IO System Model**

238 The reactive execution model above has the strict requirement that for every input, the system  
 239 produces exactly one output. Recent work [18, 19] instead considers mixed-IO models, where  
 240 a program  $L \subseteq (In \cup Out)^\omega$  is a subset of traces containing both inputs and outputs, but without  
 241 any restriction on the order or frequency in which inputs and outputs appear in the trace. In

particular, they are not required to strictly alternate (but they may, and in this way the reactive execution model can be considered a special case [16]). A particularity of this model is the distinct output symbol  $\delta$  for quiescence, i.e., the absence of an output. For example, finite behaviour can be expressed by adding infinitely many  $\delta$  symbols to a finite trace.

The new system model induces consequences regarding cleanliness contracts. Every mixed-IO trace is projected into an input, respectively output domain. The set of input symbols contains one additional element  $\neg_i$ , that indicates that in the respective steps an output was produced, but masking the concrete output. Similarly, the set of output symbols contains the additional element  $\neg_o$  to mask a concrete input symbol. *Projection on inputs*  $\downarrow_i$ :  $(\text{In} \cup \text{Out})^\omega \rightarrow (\text{In} \cup \{\neg_i\})^\omega$  and *projection on outputs*  $\downarrow_o$ :  $(\text{Out} \cup \text{Out})^\omega \rightarrow (\text{Out} \cup \{\neg_o\})^\omega$  are defined for all traces  $\sigma \in (\text{In} \cup \text{Out})^\omega$  and  $k \in \mathbb{N}$  as follows:  $\sigma \downarrow_i[k] := \text{if } \sigma[k] \in \text{In} \text{ then } \sigma[k] \text{ else } \neg_i$  and similarly  $\sigma \downarrow_o[k] := \text{if } \sigma[k] \in \text{Out} \text{ then } \sigma[k] \text{ else } \neg_o$ . The distance functions  $d_{\text{In}}$  and  $d_{\text{Out}}$  apply on input and output symbols or their respective masks, i.e., they are functions  $(\text{In} \cup \{\neg_i\}) \times (\text{In} \cup \{\neg_i\}) \rightarrow \mathbb{R}_{\geq 0}$  and, respectively,  $(\text{Out} \cup \{\neg_o\}) \times (\text{Out} \cup \{\neg_o\}) \rightarrow \mathbb{R}_{\geq 0}$ . Finally, instead of a set of standard inputs  $\text{Std}_{\text{In}}$ , we evaluate mixed-IO system cleanliness w.r.t. to a set of standard behaviour  $\text{Std} \subseteq L$ . Thus, not only inputs, but also outputs can be defined as standard behaviour and for an input, one of its outputs can be considered in  $\text{Std}$  while a different output can be excluded from  $\text{Std}$ . As a consequence, the set  $\text{Std}$  is specific for some mixed-IO system  $L$ , because  $\text{Std}$  is useful only if  $\text{Std} \subseteq L$ . To emphasise this difference we will call the tuple  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  (cleanliness) context (instead of cleanliness contract). Robust cleanliness of mixed-IO systems w.r.t. such a context is defined below [19].

**Definition 4** A mixed-IO system  $L \subseteq (\text{In} \cup \text{Out})^\omega$  is *robustly clean* w.r.t. context  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  if and only if  $\text{Std} \subseteq L$  and for all  $\sigma \in \text{Std}$  and  $\sigma' \in L$ ,

- 266 1. there exists  $\sigma'' \in L$  with  $\sigma' \downarrow_i = \sigma'' \downarrow_i$ , such that for every index  $k \in \mathbb{N}$  it holds that  
267 whenever  $d_{\text{In}}(\sigma[j] \downarrow_i, \sigma'[j] \downarrow_i) \leq \kappa_i$  for all  $j \leq k$ , then  $d_{\text{Out}}(\sigma[k] \downarrow_o, \sigma''[k] \downarrow_o) \leq \kappa_o$ ,  
268 *(l-robust cleanliness)*
- 269 2. there exists  $\sigma'' \in \text{Std}$  with  $\sigma \downarrow_i = \sigma'' \downarrow_i$ , such that for every index  $k \in \mathbb{N}$  it holds that  
270 whenever  $d_{\text{In}}(\sigma[j] \downarrow_i, \sigma'[j] \downarrow_i) \leq \kappa_i$  for all  $j \leq k$ , then  $d_{\text{Out}}(\sigma'[k] \downarrow_o, \sigma''[k] \downarrow_o) \leq \kappa_o$ .  
271 *(u-robust cleanliness)*

We will in the following refer to Definition 4.1 for l-robust cleanliness and Definition 4.2 for u-robust cleanliness. Definition 4 universally quantifies a standard trace  $\sigma$ . For l-robust cleanliness, the universal quantification of  $\sigma'$  effectively only quantifies an input sequence; the input projection for the existentially quantified  $\sigma''$  must match the projection for  $\sigma'$ . The remaining parts of the definition are conceptually identical to their reactive systems counterpart in Definition 1.1. For u-robust cleanliness, the existentially quantified trace  $\sigma''$  is obtained from set  $\text{Std}$  in contrast to l-robust cleanliness, where  $\sigma''$  can be any arbitrary trace of  $L$ . This is necessary, because u-robust cleanliness is defined w.r.t. a cleanliness context; from knowing that  $\sigma \in \text{Std}$  is a standard trace and by enforcing that  $\sigma \downarrow_i = \sigma'' \downarrow_i$  we cannot conclude that also  $\sigma'' \in \text{Std}$ .

Definition 5 shows the definition func-cleanliness of mixed-IO systems.

**Definition 5** A mixed-IO system  $L \subseteq (\text{In} \cup \text{Out})^\omega$  is *func-clean* w.r.t. context  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  if and only if  $\text{Std} \subseteq L$  and for all  $\sigma \in \text{Std}$  and  $\sigma' \in L$ ,

- 285 1. there exists  $\sigma'' \in L$  with  $\sigma' \downarrow_i = \sigma'' \downarrow_i$ , such that for every index  $k \in \mathbb{N}$ , it holds that  
286  $d_{\text{Out}}(\sigma[k] \downarrow_o, \sigma''[k] \downarrow_o) \leq f(d_{\text{In}}(\sigma[k] \downarrow_i, \sigma'[k] \downarrow_i))$ ,  
287 *(l-func-cleanliness)*

- 287 2. there exists  $\sigma'' \in \text{Std}$  with  $\sigma|_i = \sigma''|_i$ , such that for every index  $k \in \mathbb{N}$ , it holds that  
 288    $d_{\text{Out}}(\sigma'[k]|_o, \sigma''[k]|_o) \leq f(d_{\text{In}}(\sigma[k]|_i, \sigma'[k]|_i))$ . (u-func-cleanliness)

289   We will in the following refer to Definition 5.1 for l-func-cleanliness and Definition 5.2 for  
 290 u-func-cleanliness.

## 291 2.2 Temporal logics

### 292 2.2.1 HyperLTL

293 Linear Temporal Logic (LTL) [97] is a popular formalism to reason about properties of traces.  
 294 A trace is an infinite word where each literal is a subset of AP, the set of atomic propositions.  
 295 We interpret programs as circuits encoded as sets  $C \subseteq (2^{\text{AP}})^\omega$  of such traces. LTL provides  
 296 expressive means to characterise sets of traces, often called *trace properties*. For some set  
 297 of traces  $T$ , a trace property defines a subset of  $T$  (for which the property holds), whereas a  
 298 *hyperproperty* defines a *set of* subsets of  $T$  (constituting combinations of traces for which  
 299 the property holds). In this way it specifies which traces are valid in combination with one  
 300 another. Many temporal logics have been extended to corresponding hyperlogics supporting  
 301 the specification of hyperproperties.

302 HyperLTL [31] is such a temporal logic for the specification of hyperproperties of reactive  
 303 systems. It extends LTL with trace quantifiers and trace variables that make it possible to refer  
 304 to multiple traces within a logical formula. A *HyperLTL formula* is defined by the following  
 305 grammar, where  $\pi$  is drawn from a set  $\mathcal{V}$  of *trace variables* and  $a$  from the set AP:

$$\begin{aligned} \psi ::= & \exists \pi. \psi \mid \forall \pi. \psi \mid \phi \\ \phi ::= & a_\pi \mid \neg \phi \mid \phi \wedge \phi \mid X \phi \mid \phi U \phi \end{aligned}$$

306 The quantifiers  $\exists$  and  $\forall$  quantify existentially and universally, respectively, over the set of  
 307 traces. For example, the formula  $\forall \pi. \exists \pi'. \phi$  means that for every trace  $\pi$  there exists another  
 308 trace  $\pi'$  such that  $\phi$  holds over the pair of traces. To account for distinct valuations of atomic  
 309 propositions across distinct traces, the atomic propositions are indexed with trace variables:  
 310 for some atomic proposition  $a \in \text{AP}$  and some trace variable  $\pi \in \mathcal{V}$ ,  $a_\pi$  states that  $a$  holds in  
 311 the initial position of trace  $\pi$ . The temporal operators and Boolean connectives are interpreted  
 312 as usual for LTL. Further operators are derivable:  $\Diamond \phi \equiv \text{true} U \phi$  enforces  $\phi$  to eventually  
 313 hold in the future,  $\Box \phi \equiv \neg \Diamond \neg \phi$  enforces  $\phi$  to always hold, and the weak-until operator  
 314  $\phi W \phi' \equiv \phi U \phi' \vee \Box \phi$  allows  $\phi$  to always hold as an alternative to the obligation for  $\phi'$  to  
 315 eventually hold.

#### 316 *HyperLTL Characterisations of Cleanliness*

317 D'Argenio et al. [32] assume distinct sets of atomic propositions to encode inputs and outputs.  
 318 That is, they assume that  $\text{AP} = \text{AP}_i \cup \text{AP}_o$  of atomic propositions, where  $\text{AP}_i$  and  $\text{AP}_o$  are the  
 319 atomic propositions that define the the input values and, respectively, the output values. Thus,  
 320 in the context of Boolean circuit encodings of programs, we take  $\text{In} = 2^{\text{AP}_i}$  and  $\text{Out} = 2^{\text{AP}_o}$ .  
 321 We capture the following natural correspondence between reactive programs and Boolean  
 322 circuits; a circuit  $C$  can be interpreted as a function  $\hat{S} : \text{In}^\omega \rightarrow 2^{(\text{Out}^\omega)}$ , where  
 323

$$324 t \in C \quad \text{if and only if} \quad (t|_{\text{AP}_o}) \in \hat{S}(t|_{\text{AP}_i}), \quad (1)$$

325 with  $t|_A$  defined by  $(t|_A)[k] = t[k] \cap A$  for all  $k \in \mathbb{N}$ .

326   In the HyperLTL formulas below occur, for convenience, non-atomic propositions. Their  
 327 semantics is encoded by atomic propositions and Boolean connectives according to a Boolean

encoding of inputs and outputs. We refer to the original work for the details [32, Table 1]. Further, we assume that there is a quantifier-free HyperLTL formula  $\text{StdIn}_\pi$  that can check whether the trace represented by trace variable  $\pi$  is in the set of standard inputs  $\text{StdIn} \subseteq \text{In}^\omega$ . That is,  $\text{StdIn}_\pi$  should be defined such that for every trace  $t \in \mathcal{C}$  it holds that  $\{\pi := t\} \models_C \text{StdIn}_\pi$  if and only if  $(t \downarrow_{AP_i}) \in \text{StdIn}$ .

Proposition 1 shows HyperLTL formulas for l-robust cleanliness and u-robust cleanliness, respectively.<sup>1</sup>

**Proposition 1** *Let  $\mathcal{C}$  be a set of infinite traces over  $2^{\text{AP}}$ , let  $\hat{S}$  be the reactive system constructed from  $\mathcal{C}$  according to Equation 1, and let  $\mathcal{C} = \langle \text{StdIn}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  be a contract for robust cleanliness. Then  $\hat{S}$  is l-robustly clean w.r.t.  $\mathcal{C}$  if and only if  $\mathcal{C}$  satisfies the HyperLTL formula*

$$\forall \pi_1. \forall \pi_2. \exists \pi'_2. \text{StdIn}_{\pi_1} \rightarrow \left( \square(i_{\pi_2} = i_{\pi'_2}) \wedge ((d_{\text{Out}}(o_{\pi_1}, o_{\pi'_2}) \leq \kappa_o) \mathcal{W}(d_{\text{In}}(i_{\pi_1}, i_{\pi'_2}) > \kappa_i)) \right),$$

and  $\hat{S}$  is u-robustly clean w.r.t.  $\mathcal{C}$  if and only if  $\mathcal{C}$  satisfies the HyperLTL formula

$$\forall \pi_1. \forall \pi_2. \exists \pi'_1. \text{StdIn}_{\pi_1} \rightarrow \left( \square(i_{\pi_1} = i_{\pi'_1}) \wedge ((d_{\text{Out}}(o_{\pi'_1}, o_{\pi_2}) \leq \kappa_o) \mathcal{W}(d_{\text{In}}(i_{\pi'_1}, i_{\pi_2}) > \kappa_i)) \right).$$

The first quantifier (for  $\pi_1$ ) in both formulas implicitly quantifies the standard input  $i$  and the second quantifier (for  $\pi'_2$ ) implicitly quantifies the second input  $i'$ . Due to the potential nondeterminism in the behaviour of the system, the third, existential, quantifier for  $\pi'_1$ , respectively  $\pi'_2$  is necessary. While the formula for l-robust cleanliness has the universal quantification on the outputs of the program that takes the standard input  $i$  and the existential quantification on the output for  $i'$ , the formula for u-robust cleanliness works in the other way around. Thus, the formulas capture the  $\forall \exists$  alternation in Definition 1. The weak until operator  $\mathcal{W}$  has exactly the behaviour necessary to represent the interaction between the distances of inputs and the distances of outputs.

The HyperLTL formulas for func-cleanliness are given below.

**Proposition 2** *Let  $\mathcal{C}$  be a set of infinite traces over  $2^{\text{AP}}$ , let  $\hat{S}$  be the reactive system constructed from  $\mathcal{C}$  according to Equation 1, and let  $\mathcal{C} = \langle \text{StdIn}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  be a contract for func-cleanliness. Then  $\hat{S}$  is l-func-clean w.r.t.  $\mathcal{C}$  if and only if  $\mathcal{C}$  satisfies the HyperLTL formula*

$$\forall \pi_1. \forall \pi_2. \exists \pi'_2. \text{StdIn}_{\pi_1} \rightarrow \left( \square(i_{\pi_2} = i_{\pi'_2}) \wedge \square(d_{\text{Out}}(o_{\pi_1}, o_{\pi'_2}) \leq f(d_{\text{In}}(i_{\pi_1}, i_{\pi'_2}))) \right),$$

and  $\hat{S}$  is u-func-clean w.r.t.  $\mathcal{C}$  if and only if  $\mathcal{C}$  satisfies the HyperLTL formula

$$\forall \pi_1. \forall \pi_2. \exists \pi'_1. \text{StdIn}_{\pi_1} \rightarrow \left( \square(i_{\pi_1} = i_{\pi'_1}) \wedge \square(d_{\text{Out}}(o_{\pi'_1}, o_{\pi_2}) \leq f(d_{\text{In}}(i_{\pi'_1}, i_{\pi_2}))) \right).$$

### 2.2.2 Signal temporal logic

LTL enables reasoning over traces  $\sigma \in (2^{\text{AP}})^\omega$  for which it is necessary to encode values using the atomic propositions in AP. Each literal in a trace represents a discrete time step of an underlying model. Thus,  $\sigma$  can equivalently be viewed as a function  $\mathbb{N} \rightarrow 2^{\text{AP}}$ . One extension of LTL is *Signal Temporal Logic* (STL) [33, 76], which instead is used for reasoning over

<sup>1</sup> All HyperLTL formulas from D'Argenio et al. [32] are adapted for non-parametrised systems.

real-valued signals that may change in value along an underlying continuous time domain. In this article, we generalise the original work and use *generalised timed traces* (GTTs) [53], which, for some value domain  $X$  and time domain  $\mathcal{T}$  define traces as functions  $\mathcal{T} \rightarrow X$ . The time domain  $\mathcal{T}$  can be either  $\mathbb{N}$  (*discrete-time*), or  $\mathbb{R}_{\geq 0}$  (*continuous-time*). For the value domain we will use vectors of real values  $X = \mathbb{R}^n$  for some  $n > 0$  or, to express mixed-IO traces, the set  $X = \text{In} \cup \text{Out}$ .

STL formulas can express properties of systems modelled as sets  $M \subseteq (\mathcal{T} \rightarrow X)$  of traces by making the atomic properties refer to booleanisations of the signal values. The syntax of the variant of STL that we use in this article is as follows, where  $f \in X \rightarrow \mathbb{R}$ :

$$\phi ::= \top \mid f > 0 \mid \neg\phi \mid \phi \wedge \psi \mid \phi \mathcal{U} \psi.$$

STL replaces atomic propositions by *threshold predicates* of the form  $f > 0$ , which hold if and only if function  $f$  applied to the trace value at the current time returns a positive value. The Boolean operators and the Until operator  $\mathcal{U}$  are very similar to those of HyperLTL. The Next operator  $\mathcal{X}$  is not part of STL, because “next” is without precise meaning in continuous time. The definitions of the derived operators  $\diamondsuit$ ,  $\square$  and  $\mathcal{W}$  are the same as for HyperLTL. Formally, the *Boolean semantics* of an STL formula  $\phi$  at time  $t \in \mathcal{T}$  for a trace  $w \in \mathcal{T} \rightarrow X$  is defined inductively:

$$\begin{aligned} w, t \models \top \\ w, t \models f > 0 &\quad \text{iff } f(w(t)) > 0 \\ w, t \models \neg\phi &\quad \text{iff } w, t \not\models \phi \\ w, t \models \phi \wedge \psi &\quad \text{iff } w, t \models \phi \text{ and } w, t \models \psi \\ w, t \models \phi \mathcal{U} \psi &\quad \text{iff exists } t' \geq t \text{ s.t. } w, t' \models \psi \text{ and} \\ &\quad \text{for all } t'' \in [t, t'), w, t'' \models \phi \end{aligned}$$

A system  $M$  satisfies a formula  $\phi$ , denoted  $M \models \phi$ , if and only if for every  $w \in M$  it holds that  $w, 0 \models \phi$ .

### Quantitative Interpretation

STL has been extended by a *quantitative semantics* [1, 33, 49]. This semantics is designed in such a way that whenever  $\rho(\phi, w, t) \neq 0$ , its sign indicates whether  $w, t \models \phi$  holds in the Boolean semantics. For any STL formula  $\phi$ , trace  $w$  and time  $t$ , if  $\rho(\phi, w, t) > 0$ , then  $w, t \models \phi$  holds, and if  $\rho(\phi, w, t) < 0$ , then  $w, t \models \phi$  does not hold. The quantitative semantics for an STL formula  $\phi$ , trace  $w$ , and time  $t$  the quantitative semantics is defined inductively:

$$\begin{aligned} \rho(\top, w, t) &= \infty \\ \rho(f > 0, w, t) &= f(w(t)) \\ \rho(\neg\phi, w, t) &= -\rho(\phi, w, t) \\ \rho(\phi \wedge \psi, w, t) &= \min(\rho(\phi, w, t), \rho(\psi, w, t)) \\ \rho(\phi \mathcal{U} \psi, w, t) &= \sup_{t' \geq t} \min\{\rho(\psi, w, t'), \inf_{t'' \in [t, t')} \rho(\phi, w, t'')\} \end{aligned}$$

### Robustness and Falsification

The value of the quantitative semantics can serve as a *robustness estimate* and as such be used to search for a violation of the property at hand, i.e., to falsify it. The robustness of STL formula  $\phi$  is its quantitative value at time 0, that is,  $\mathcal{R}_\phi(w) := \rho(\phi, w, 0)$ . So, falsifying a formula  $\phi$  for a system  $M$  boils down to a search problem with the goal condition  $\mathcal{R}_\phi(w) < 0$ .

**Algorithm 1** Monte-Carlo falsification

**Input:**  $w$ : Initial trace,  $\mathcal{R}$ : Robustness function, PS: Proposal Scheme

**Output:**  $w \in M$

```

1: while  $\mathcal{R}(w) > 0$  do
2:    $w' \leftarrow \text{PS}(w)$ 
3:    $\alpha \leftarrow \exp(-\beta(\mathcal{R}(w') - \mathcal{R}(w)))$ 
4:    $r \leftarrow \text{UniformRandomReal}(0, 1)$ 
5:   if  $r \leq \alpha$  then
6:      $w \leftarrow w'$ 
7:   end if
8: end while

```

Successful falsification algorithms solve this problem by understanding it as the optimisation problem  $\text{minimise}_{w \in M} \mathcal{R}_\phi(w)$ . Algorithm 1 [1, 88] sketches an algorithm for Monte-Carlo Markov Chain falsification, which is based on acceptance-rejection sampling [29].

An input to the algorithm is an initial trace  $w$  and a computable robustness function  $\mathcal{R}$ . Robustness computation for STL formulas has been addressed in the literature [33, 49]; we omit this discussion here. The third input PS is a proposal scheme that proposes a new trace to the algorithm based on the previous one (line 2). The parameter  $\beta$  (used in line 3) can be adjusted during the search and is a means to avoid being trapped in local minima, preventing to find a global minimum.

Notably, there exists prior work by Nguyen et al. [89] that discusses an extension of STL to HyperSTL though using a non-standard semantic underpinning. In this context, they present a falsification approach restricted to the fragment “t-HyperSTL” where, according to the authors, “a nesting structure of temporal logic formulas involving different traces is not allowed”. Therefore, none of our cleanliness definitions belongs to this fragment.

### 3 Logical characterisation of Mixed-IO cleanliness

In this section we provide a temporal logic characterisation for robust cleanliness and func-cleanliness for mixed-IO systems. For this, we propose a HyperSTL semantics (different to that of [89]) and propose HyperSTL formulas for robust cleanliness and func-cleanliness. We explain how these formulas can be applied to mixed-IO traces and prove that the characterisation is correct. Furthermore, for the special case that Std is a finite set, we reformulate the HyperSTL formulas characterising the u-cleanlinesses as equivalent STL formulas.

#### *Hyperlogics over Continuous Domains*

Previous work [89] extends STL to HyperSTL echoing the extension of LTL to HyperLTL. We use a similar HyperSTL syntax in this article:

$$\begin{aligned}\psi ::= & \exists \pi. \psi \mid \forall \pi. \psi \mid \phi \\ \phi ::= & \top \mid f > 0 \mid \neg \phi \mid \phi \wedge \phi \mid \phi \mathcal{U} \phi.\end{aligned}$$

The meaning of the universal and existential quantifier is as for HyperLTL. In contrast to HyperLTL (and to the existing definition of HyperSTL), we consider it insufficient to allow propositions to refer to only a single trace. In HyperLTL atomic propositions of individual traces can be compared by means of the Boolean connectives. To formulate thresholds for real values, however, we feel the need to allow real values from multiple traces to be combined in the function  $f$ , and thus to appear as arguments of  $f$ . Hence, in our semantics of HyperSTL,  $f > 0$  holds if and only if the result of  $f$ , applied to all traces quantified over, is greater than

438 0. For this to work formally, the arity of function  $f$  is the number  $m$  of traces quantified over  
 439 at the occurrence of  $f > 0$  in the formula, so  $f : X^m \rightarrow \mathbb{R}$ .

440 A trace assignment [31]  $\Pi : \mathcal{V} \rightarrow M$  is a partial function assigning traces of  $M$  to variables.  
 441 Let  $\Pi[\pi := w]$  denote the same function as  $\Pi$ , except that  $\pi$  is mapped to trace  $w$ . The Boolean  
 442 semantics of HyperSTL is defined below.

443 **Definition 6** Let  $\psi$  be a HyperSTL formula,  $t \in \mathcal{T}$  a time point,  $M \subseteq (\mathcal{T} \rightarrow X)$  a set of  
 444 GTTs, and  $\Pi$  a trace assignment. Then, the Boolean semantics for  $M, \Pi, t \models \psi$  is defined  
 445 inductively:

$$\begin{aligned} 446 \quad M, \Pi, t \models \exists \pi. \psi &\Leftrightarrow \exists w \in M. M, \Pi[\pi := w], t \models \psi \\ 447 \quad M, \Pi, t \models \forall \pi. \psi &\Leftrightarrow \forall w \in M. M, \Pi[\pi := w], t \models \psi \\ 448 \quad M, \Pi, t \models \top & \\ 449 \quad M, \Pi, t \models f > 0 &\Leftrightarrow f(\Pi(\pi_1)(t), \dots, \Pi(\pi_m)(t)) > 0 \text{ for } \text{dom}(\Pi) = \{\pi_1, \dots, \pi_m\}^2 \\ 450 \quad M, \Pi, t \models \neg \phi &\Leftrightarrow M, \Pi, t \not\models \phi \\ 451 \quad M, \Pi, t \models \phi_1 \wedge \phi_2 &\Leftrightarrow M, \Pi, t \models \phi_1 \text{ and } M, \Pi, t \models \phi_2 \\ 452 \quad M, \Pi, t \models \phi_1 \cup \phi_2 &\Leftrightarrow \exists t' \geq t. M, \Pi, t' \models \phi_2 \text{ and } \forall t'' \in [t, t'). M, \Pi, t'' \models \phi_1 \end{aligned}$$

453 A system  $M$  satisfies a formula  $\psi$  if and only if  $M, \emptyset, 0 \models \psi$ . The quantitative semantics  
 454 for HyperSTL is defined below:

455 **Definition 7** Let  $\psi$  be a HyperSTL formula,  $t \in \mathcal{T}$  a time point,  $M \subseteq (\mathcal{T} \rightarrow X)$  a set of  
 456 GTTs, and  $\Pi$  a trace assignment. Then, the quantitative semantics for  $\rho(\psi, M, \Pi, t)$  is  
 457 defined inductively:

$$\begin{aligned} 458 \quad \rho(\exists \pi. \psi, M, \Pi, t) &= \sup_{w \in M} \rho(\psi, M, \Pi[\pi := w], t) \\ 459 \quad \rho(\forall \pi. \psi, M, \Pi, t) &= \inf_{w \in M} \rho(\psi, M, \Pi[\pi := w], t) \\ 460 \quad \rho(\top, M, \Pi, t) &= \infty \\ 461 \quad \rho(f > 0, M, \Pi, t) &= f(\Pi(\pi_1)(t), \dots, \Pi(\pi_m)(t)) \text{ for } \text{dom}(\Pi) = \{\pi_1, \dots, \pi_m\}^2 \\ 462 \quad \rho(\neg \phi, M, \Pi, t) &= -\rho(\phi, M, \Pi, t) \\ 463 \quad \rho(\phi_1 \wedge \phi_2, M, \Pi, t) &= \min(\rho(\phi_1, M, \Pi, t), \rho(\phi_2, M, \Pi, t)) \\ 464 \quad \rho(\phi_1 \cup \phi_2, M, \Pi, t) &= \sup_{t' \geq t} \min\{\rho(\phi_2, M, \Pi, t'), \inf_{t'' \in [t, t')} \rho(\phi_1, M, \Pi, t'')\} \end{aligned}$$

### 465 *HyperSTL Characterisation*

466 The HyperLTL characterisations in Sect. 2.2.1 assume the system to be a subset of  $(2^{AP})^\omega$   
 467 and works with distances between traces by means of a Boolean encoding into atomic propo-  
 468 sitions. By using HyperSTL, we can characterise cleanliness for systems that are representable  
 469 as subsets of  $(\mathcal{T} \rightarrow X)$ .

470 We can take the HyperLTL formulas from Proposition 1 and 2 and transform them into  
 471 HyperSTL formulas by applying simple syntactic changes. We get for l-robust cleanliness the  
 472 formula

$$\begin{aligned} 473 \quad \psi_{l\text{-rob}} &:= \forall \pi_1. \forall \pi_2. \exists \pi'_2. \text{Std}_{\pi_1} > 0 \\ 474 \quad &\rightarrow \left( \square(\text{eq}(\pi_2 \downarrow_i, \pi'_2 \downarrow_i) \leq 0) \wedge \right. \end{aligned}$$

<sup>2</sup> We admit some sloppiness; the set  $\text{dom}(\Pi)$  should have a fixed order.

$$(d_{\text{Out}}(\pi_1 \downarrow_o, \pi'_1 \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(\pi_1 \downarrow_i, \pi'_1 \downarrow_i) - \kappa_i > 0)), \quad (2)$$

476 u-robust cleanliness is characterised by

$$\begin{aligned} 477 \quad \psi_{\text{u-rob}} := & \forall \pi_1. \forall \pi_2. \exists \pi'_1. \text{Std}_{\pi_1} > 0 \\ 478 \quad & \rightarrow (\text{Std}_{\pi'_1} > 0 \wedge \square(\text{eq}(\pi_1 \downarrow_i, \pi'_1 \downarrow_i) \leq 0) \wedge \\ 479 \quad & ((d_{\text{Out}}(\pi'_1 \downarrow_o, \pi_2 \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(\pi'_1 \downarrow_i, \pi_2 \downarrow_i) - \kappa_i > 0))), \end{aligned} \quad (3)$$

480 for l-func-cleanliness we get the formula

$$\begin{aligned} 481 \quad \psi_{\text{l-fun}} := & \forall \pi_1. \forall \pi_2. \exists \pi'_2. \text{Std}_{\pi_1} > 0 \\ 482 \quad & \rightarrow (\square(\text{eq}(\pi_2 \downarrow_i, \pi'_2 \downarrow_i) \leq 0) \wedge (\square(d_{\text{Out}}(\pi_1 \downarrow_o, \pi'_2 \downarrow_o) - f(d_{\text{In}}(\pi_1 \downarrow_i, \pi'_2 \downarrow_i)) \leq 0))), \end{aligned} \quad (4)$$

483 and, finally, u-func-cleanliness is encoded by

$$\begin{aligned} 484 \quad \psi_{\text{u-fun}} := & \forall \pi_1. \forall \pi_2. \exists \pi'_1. \text{Std}_{\pi_1} > 0 \\ 485 \quad & \rightarrow (\text{Std}_{\pi'_1} > 0 \wedge \square(\text{eq}(\pi_1 \downarrow_i, \pi'_1 \downarrow_i) \leq 0) \wedge \\ 486 \quad & ((d_{\text{Out}}(\pi'_1 \downarrow_o, \pi_2 \downarrow_o) - f(d_{\text{In}}(\pi'_1 \downarrow_i, \pi_2 \downarrow_i)) \leq 0))). \end{aligned} \quad (5)$$

487 The quantifiers remain unchanged relative to the formulas in Propositions 1 and 2. The  
488 formulas use generic projection functions  $\downarrow_i : X \rightarrow \text{In}$  and  $\downarrow_o : X \rightarrow \text{Out}$  to extract the input  
489 values, respectively output values from a trace. To apply the formulas, these functions must be  
490 instantiated with functions for the concrete instantiation of the value domain  $X$  of the traces to  
491 be analysed. For example, for  $\text{In} = \mathbb{R}^m$ ,  $\text{Out} = \mathbb{R}^l$ , and  $M \subseteq (\mathcal{T} \rightarrow \mathbb{R}^{m+l})$ , the projections  
492 could be defined for every  $w = (s_1, \dots, s_m, s_{m+1}, \dots, s_{m+l})$  as  $w \downarrow_i = (s_1, \dots, s_m)$  and  
493  $w \downarrow_o = (s_{m+1}, \dots, s_{m+l})$ . The input equality requirement for two traces  $\pi$  and  $\pi'$  is ensured  
494 by globally enforcing  $\text{eq}(\pi \downarrow_i, \pi' \downarrow_i) \leq 0$ .  $\text{eq}$  is a generic function that returns zero if its  
495 arguments are identical and a positive value otherwise. It must be instantiated for concrete  
496 value domains. For example,  $\text{eq}((s_1, \dots, s_m), (s'_1, \dots, s'_m))$  could be defined as the sum of  
497 the component-wise distances  $\sum_{1 \leq i \leq m} |s_i - s'_i|$ . Finally, in the above formulas we perform  
498 simple arithmetic operations to match the syntactic requirements of HyperSTL.

499 Formulas (3) and (5) are prepared to express u-robust cleanliness, respectively u-func-  
500 cleanliness w.r.t. both cleanliness *contracts* or cleanliness *contexts*. That is, we assume the  
501 existence of a function  $\text{Std}_\pi$  that returns a positive value if and only if the trace assigned to  
502  $\pi$  encodes a standard input (when considering cleanliness *contracts*) or encodes an input and  
503 output that constitute a standard behaviour (when considering cleanliness *contexts*). Explicitly  
504 requiring that  $\pi'_1$  represents a standard behaviour echoes the setup in Definitions 4.2 and 5.2.

505 We remark that for encoding  $\text{Std}_\pi$ , due to the absence of the Next-operator in HyperSTL,  
506 it might be necessary to add a clock signal  $s(t) = t$  to traces in a preprocessing step.

507 **Example 3** Let  $\text{In} = \text{Out} = \mathbb{R}$  be the sets representing real-valued inputs and outputs,  $\mathcal{T} = \mathbb{N}$   
508 be the discrete time domain, and  $X = \text{In} \times \text{Out}$  the value domain that considers pairs of inputs  
509 and outputs as values. We consider the robust cleanliness context  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$ ,  
510 where  $\text{Std} = \{w_0, w_1\}$  contains the two standard traces

$$511 \quad w_0 = (1; 0) (2; 0) (3; 0) (4; 0) \dots \text{ and } w_1 = (1; 1) (2; 2) (3; 3) (4; 4) \dots$$

512 For the distance functions we use the absolute differences, i.e.,  $d_{\text{In}}(v_1, v_2) = d_{\text{Out}}(v_1, v_2) =$   
513  $|v_1 - v_2|$ . Let the value thresholds be  $\kappa_i = 1$  and  $\kappa_o = 2$ , and let  $\downarrow_i, \downarrow_o, \text{eq}$   
514 and  $\text{Std}_\pi$  be defined as explained above. We consider the non-standard traces  $w_A =$   
515  $(1.3; 0) (2.6; 0) (3.9; 0) (5.2; 0) \dots$ ,  $w_B = (1.3; 1.3) (2.6; 2.6) (3.9; 3.9) (5.2; 5.2) \dots$ ,  
516 and  $w_x = (1.5; 1.5) (2.5; 3.2) (3.5; 4.9) (4.5; 6.6) \dots$ .

The HyperSTL formulas  $\psi_{l\text{-rob}}$  and  $\psi_{u\text{-rob}}$  reason about sets of traces. For example, the set  $M = \{w_0, w_1, w_A, w_B\}$  satisfies both formulas. If both  $\pi_1$  and  $\pi_2$  represent standard traces, then  $\pi_1 \downarrow_i = \pi_2 \downarrow_i$ , because  $w_0 \downarrow_i = w_1 \downarrow_i$ , and the formulas hold for  $\pi'_2 = \pi_1$ , respectively  $\pi'_1 = \pi_2$ . Otherwise, assume that  $\pi_1$  represents  $w_0$  and  $\pi_2$  represents  $w_B$  (the reasoning for other combinations of traces is similar).

First considering  $\psi_{l\text{-rob}}$ , we pick  $w_A$  for  $\pi'_2$ . We get that  $\pi_2 \downarrow_i = \pi'_2 \downarrow_i$ , because  $w_B \downarrow_i = w_A \downarrow_i$ . Hence, we globally have  $|\pi_2 \downarrow_i - \pi'_2 \downarrow_i| = 0$  and, thus,  $\text{eq}(\pi_2 \downarrow_i, \pi'_2 \downarrow_i) = 0$ . At time steps  $0 \leq t \leq 3$ , the distance between the outputs  $|w_0 \downarrow_o(t) - w_A \downarrow_o(t)|$  is at most  $\kappa_o$ . Hence, the left operand of  $\mathcal{W}$  holds and the formula is satisfied for  $t \leq 3$ . At time  $t = 3$  we have that  $|w_0 \downarrow_i(t) - w_A \downarrow_i(t)| = |4.0 - 5.2| > \kappa_i$ . Hence, the right operand of the  $\mathcal{W}$  operator holds and  $\psi_{l\text{-rob}}$  is satisfied also for  $t \geq 3$ . Notice that if we would remove  $w_A$  from  $M$ , then it would violate  $\psi_{l\text{-rob}}$ , because there is no possible choice for  $\pi'_2$  that has the same inputs as  $w_B$  and where the output distances to  $w_0$  are below the  $\kappa_o$  threshold.

To satisfy  $\psi_{u\text{-rob}}$ , we pick  $w_1$  for  $\pi'_1$ . The reasoning why the formula holds for this choice is analogue to  $\psi_{l\text{-rob}}$ . Notice that if we add the trace  $w_x$  to  $M$ , then  $\psi_{u\text{-rob}}$  is violated. Concretely,  $\pi_2$  could represent  $w_x$ ; then, whether we pick  $w_0$  or  $w_1$  for  $\pi'_1$ , we eventually get outputs that violate the  $\kappa_o$  constraint, while the  $\kappa_i$  constraint is always satisfied. For example, if we compare  $w_x$  and  $w_1$ , then we have for all time steps  $t \leq 3$  that  $|w_1 \downarrow_i(t) - w_x \downarrow_i(t)| = 0.5 \leq \kappa_i$ , but at time  $t = 3$  we get  $|w_1 \downarrow_o(t) - w_x \downarrow_o(t)| = 2.6 > \kappa_o$ . Hence, at  $t = 3$  the left and right operand of  $\mathcal{W}$  are false, so  $\psi_{u\text{-rob}}$  is violated.

### Correctness under Mixed-IO Interpretation

Mixed-IO signals are defined in the discrete time domain  $\mathbb{N}$  and value domain  $\text{In} \cup \text{Out}$ . The abstract functions  $\downarrow_i$  and  $\downarrow_o$  can be defined equally to the syntactically identical projection functions for mixed-IO models defined in Sect. 2.1. The function  $\text{eq}(i_1, i_2)$  can be defined using the distance function  $d_{\text{In}}$  and some arbitrary small  $\varepsilon > 0$ :

$$\text{eq}(i_1, i_2) := \begin{cases} 0, & \text{if } i_1 = i_2 \\ d_{\text{In}}(i_1, i_2) + \varepsilon, & \text{if } i_1 \neq i_2 \wedge i_1, i_2 \in \text{In} \\ \infty, & \text{otherwise.} \end{cases} \quad (6)$$

In the second clause of the above definition we add some positive value  $\varepsilon$  to the result of  $d_{\text{In}}$ , because  $d_{\text{In}}(i_1, i_2)$  could be 0 even if  $i_1 \neq i_2$ . For the correctness of the above HyperSTL formulas, however, it is crucial that  $\text{eq}(i_1, i_2) = 0$  if and only if  $i_1 = i_2$ . For a good performance of the falsification algorithm, we will nevertheless want to make use of  $d_{\text{In}}$  if  $i_1 \neq i_2$ .

Proposition 3 shows that HyperSTL formulas (2) and (3) under the mixed-IO interpretation outlined above indeed characterise l-robust cleanliness and u-robust cleanliness. Proposition 4 shows the same for func-cleanliness.

**Proposition 3** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  a contract or context for robust cleanliness with  $\text{Std} \subseteq L$ . Further, let  $\text{Std}_\pi$  be a quantifier-free HyperSTL subformula, such that  $L, \{\pi := w\}, 0 \models \text{Std}_\pi$  if and only if  $w \in \text{Std}$ . Then,  $L$  is l-robustly clean w.r.t.  $C$  if and only if  $L, \emptyset, 0 \models \psi_{l\text{-rob}}$ , and  $L$  is u-robustly clean w.r.t.  $C$  if and only if  $L, \emptyset, 0 \models \psi_{u\text{-rob}}$ .

**Proposition 4** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  a contract or context for func-cleanliness with  $\text{Std} \subseteq L$ . Further, let  $\text{Std}_\pi$  be a quantifier-free HyperSTL subformula, such that  $L, \{\pi := w\}, 0 \models \text{Std}_\pi$  if and only if  $w \in \text{Std}$ . Then,  $L$  is l-func-clean w.r.t.  $C$  if and only if  $L, \emptyset, 0 \models \psi_{l\text{-fun}}$ , and  $L$  is u-func-clean w.r.t.  $C$  if and only if  $L, \emptyset, 0 \models \psi_{u\text{-fun}}$ .

### 560 **STL Characterisation for Finite Standard Behaviour**

561 In many practical settings—when the different standard behaviours are spelled out upfront  
 562 explicitly, as in NEDC and WLTC—it can be assumed that the number of distinct standard  
 563 behaviours  $\text{Std}$  is finite (while there are infinitely many possible behaviours in  $M$ ). Finiteness  
 564 of  $\text{Std}$  makes it possible to remove by enumeration the quantifiers from the  $u$ -robust cleanliness  
 565 and  $u$ -func-cleanliness HyperSTL formulas. This opens the way to work with the STL fragment  
 566 of HyperSTL, after proper adjustments. In the following, we assume that the set  $\text{Std} =$   
 567  $\{w_1, \dots, w_c\}$  is an arbitrary standard set with  $c$  unique standard traces, where every  $w_k : T \rightarrow X$  uses the same time domain  $T$  and value domain  $X$ .

568 To encode the HyperSTL formulas (3) and (5) in STL, we use the concept of *self-composition*,  
 569 which has proven useful for the analysis of hyperproperties [9, 51]. We  
 570 concatenate a trace under analysis  $w : T \rightarrow X$  and the standard traces  $w_1$  to  $w_c$  to the  
 571 composed trace  $w_+ = (w, w_1, \dots, w_c) \subseteq (T \rightarrow X^{c+1})$ . Given a system  $M \subseteq (T \rightarrow X)$   
 572 and a set  $\text{Std} = \{w_1, \dots, w_c\} \subseteq M$ , we denote by  $M \circ \text{Std} := \{(w, w_1, \dots, w_c) \mid w \in M\}$   
 573 the system in which every trace in  $M$  is composed with the standard traces in  $\text{Std}$ . For every  
 574  $w_+ \in M \circ \text{Std}$ , we will in the following STL formula write  $w$  to mean the projection on  $w_+$   
 575 to the trace  $w$ , and we write  $w_k$ , for  $1 \leq k \leq c$ , to mean the projection on  $w_+$  to the  $k$ th  
 576 standard trace.

578 **Theorem 5** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  a  
 579 context for robust cleanliness with finite standard behaviour  $\text{Std} = \{w_1, \dots, w_c\} \subseteq L$ . Then,  
 580  $L$  is  $u$ -robustly clean w.r.t.  $C$  if and only if  $(L \circ \text{Std}) \models \varphi_{u\text{-rob}}$ , where

$$581 \varphi_{u\text{-rob}} := \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} \left( \square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0) \wedge \right. \\ 582 \quad \left. ((d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i) - \kappa_i > 0)) \right).$$

583 The theorem for  $u$ -func-cleanliness is analogue to Theorem 5.

584 **Theorem 6** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  a  
 585 context for func-cleanliness with finite standard behaviour  $\text{Std} = \{w_1, \dots, w_c\} \subseteq L$ . Then,  $L$   
 586 is  $u$ -func-clean w.r.t.  $C$  if and only if  $(L \circ \text{Std}) \models \varphi_{u\text{-fun}}$ , where

$$587 \varphi_{u\text{-fun}} := \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} \left( \square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0) \wedge \right. \\ 588 \quad \left. (\square(d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) - f(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i)) \leq 0)) \right).$$

589 **Example 4** We consider the robust cleanliness context  $C = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$  where  $\text{Std} =$   
 590  $\{w_1, w_2\}$  contains the two standard traces  $w_1 = 1_i 2_i 3_i 7_o 0_i \delta^\omega$  and  $w_2 = 0_i 1_i 2_i 3_i 6_o \delta^\omega$ .  
 591 We here decorate inputs with index  $i$  and outputs with index  $o$ , i.e.,  $w_1$  describes a system  
 592 receiving the three inputs 1, 2, and 3, then producing the output 7, and finally receiving input  
 593 0 before entering quiescence. We take

$$594 d_{\text{In}}(i_1, i_2) = \begin{cases} |i_1 - i_2|, & \text{if } i_1, i_2 \in \text{In} \\ 0, & \text{if } i_1 = i_2 = -i \\ \infty, & \text{otherwise,} \end{cases}$$

595 and

596

$$d_{\text{Out}}(o_1, o_2) = \begin{cases} |o_1 - o_2|, & \text{if } o_1, o_2 \in \text{Out} \setminus \{\delta\} \\ 0, & \text{if } o_1 = o_2 = -_o \text{ or } o_1 = o_2 = \delta \\ \infty, & \text{otherwise.} \end{cases}$$

597 The contractual value thresholds are assumed to be  $\kappa_i = 1$  and  $\kappa_o = 6$ .

598 Assume we are observing the trace  $w = 0_i 1_i 2_i 6_o 0_i \delta^\omega$  to be monitored with STL formula  
 599  $\varphi_{\text{u-rob}}$  (from Theorem 5). First notice, that for combinations of  $a$  and  $b$  in  $\varphi_{\text{u-rob}}$ , where  $a \neq b$ ,  
 600 the subformula  $\square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0)$  is always false, because  $w_1$  and  $w_2$  have different  
 601 (input) values at time point 0. Hence, it remains to show that

602 
$$(d_{\text{Out}}(w_1 \downarrow_o, w \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(w_1 \downarrow_i, w \downarrow_i) - \kappa_i > 0) \wedge$$
  
 603 
$$(d_{\text{Out}}(w_2 \downarrow_o, w \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(w_2 \downarrow_i, w \downarrow_i) - \kappa_i > 0).$$

604 For the first conjunct, the input distance between inputs in  $w$  and  $w_1$  is always 1 at positions 1  
 605 to 3, it is 0 at position 4 (because  $-_i$  is compared to  $-_i$ ), and remains 0 in position 5 and beyond.  
 606 Thus,  $d_{\text{In}}(w_1 \downarrow_i, w \downarrow_i) - \kappa_i$  is always at most 0, and the right hand-side of the  $\mathcal{W}$  operator  
 607 is always false. Consequently, by definition of  $\mathcal{W}$ , the left operand of  $\mathcal{W}$  must always hold,  
 608 i.e.,  $d_{\text{Out}}(w_1 \downarrow_o, w \downarrow_o)$  must always be less or equal to 6. This is the case for  $w_1$  and  $w$ : at all  
 609 positions except for 4,  $-_o$  is compared to  $-_o$  (or  $\delta$  to  $\delta$ ), so the difference is 0, and at position  
 610 4, the distance of 6 and 7 is 1.

611 For the second  $\mathcal{W}$ -formula,  $w$  is compared to  $w_2$ . These two traces are comparable only to  
 612 a limited extent: the order of input and output is altered at the last two positions of the signals  
 613 before quiescence. Hence, the right operand of  $\mathcal{W}$  is true at position 4, and the formula holds  
 614 for the remaining trace. For positions 1 to 3, the input distances are 0, because the input values  
 615 are identical. At these positions, the left operand must hold. The values are input values, so  
 616  $-_o$  is compared to  $-_o$  at each position. This distance is defined to be 0, so it holds that  $-6 \leq 0$ ,  
 617 and the formula is satisfied. Since both formulas hold, the conjunction of both holds, too,  
 618 and trace  $w$  is qualified as robustly clean. There could however be other system traces not  
 619 considered in this example, that overall could violate robust cleanliness of the system.

## 620 **Restriction of input space**

621 Robust cleanliness puts semantic requirements on fragments of a system's input space, outside  
 622 of which the system's behaviour remains unspecified. Typically, the fragment of the input  
 623 space covered is rather small. To falsify the STL formula  $\varphi_{\text{u-rob}}$  from Theorem 5, the falsifier  
 624 has two challenging tasks. First, it has to find a way to stay in the relevant input space,  
 625 i.e., select inputs with a distance of at most  $\kappa_i$  from the standard behaviour. Only if this is  
 626 assured it can search for an output large enough to violate the  $\kappa_o$  requirement. In this, a  
 627 large robustness estimate provided by the quantitative semantics of STL cannot serve as an  
 628 indicator for deciding whether an input is too far off or whether an output stays too close to the  
 629 standard behaviour. We can improve the efficiency of the falsification process significantly  
 630 by narrowing upfront the input space the falsifier uses.

631 In practice, test execution traces will always be finite. In previous real-life doping tests,  
 632 test execution lengths have been bounded by some constant  $B \in \mathbb{N}$  [19], i.e., systems are  
 633 represented as sets of finite traces  $M \subseteq (\text{In} \cup \text{Out})^B$  (which for formality reasons each  
 634 can be considered suffixed with  $\delta^\omega$ ). In this bounded horizon, we can provide a predicate  
 635 discriminating between relevant and irrelevant input sequences. Formally, the restriction  
 636 to the relevant input space fragment of a system  $M \subseteq (\text{In} \cup \text{Out})^B$  is given by the set  
 637  $\text{In}_{\text{Std}, \kappa_i} = \{w \in M \mid \exists w' \in \text{Std}. \bigwedge_{k=0}^{B-1} (d_{\text{In}}(w[k] \downarrow_i, w'[k] \downarrow_i) \leq \kappa_i)\}$ . Since  $\text{Std}$  and  $B$  are  
 638 finite, membership is computable.

639 There are rare cases in which this optimisation may prevent the falsifier from finding a  
 640 counterexample. This is only the case if there is an input prefix leading to a violation of the  
 641 formula for which there is no suffix such that the whole trace satisfies the  $\kappa_i$  constraint. Below  
 642 is a pathological example in which this could make a difference.

643 **Example 5** Apart from  $\text{NO}_x$  emissions, NEDC (and WLTC) tests are used to measure fuel  
 644 consumption. Consider a contract similar to the contracts above, but with fuel rate as the  
 645 output quantity. Assuming a “normal” fuel rate behaviour during the standard test, there  
 646 might be a test within a reasonable  $\kappa_i$  distance, where the fuel is wasted insanely. Then, the  
 647 fuel tank might run empty before the intended end of the test, which therefore could not be  
 648 finished within the  $\kappa_i$  distance, because speed would be constantly 0 at the end. The actually  
 649 driven test is not in set  $\text{In}_{\text{Std}, \kappa_i}$ , but there is a prefix within  $\kappa_i$  distance that violates the robust  
 650 cleanliness property.

651 Notably, there may be additional techniques to reduce the size of the input space. For  
 652 example, if the next input symbol depends on the history of inputs, this constraint could be  
 653 considered in the proposal scheme.

## 654 4 Supervision of diesel emission cleaning systems

655 The severity of the diesel emissions scandal showed that the regulations alone are insufficient  
 656 to prevent car manufacturers from implementing tampered—or doped—emission cleaning  
 657 systems. Recent works [19] shows that robust cleanliness is a suitable means to extend the  
 658 precisely defined behaviour of cars for the NEDC to test cycles within a  $\kappa_i$  range around  
 659 the NEDC. To demonstrate the usefulness of robust cleanliness, the essential details of the  
 660 emission testing scenario were modelled: the set of inputs is the set of speed values, an output  
 661 value represents the amount of emissions—in particular, the nitric oxide ( $\text{NO}_x$ ) emissions—  
 662 measured at the exhaust pipe of a car. The distance functions are the absolute differences  
 663 of speed, respectively  $\text{NO}_x$ , values, and the standard behaviour is the singleton set that  
 664 contains a trace that consists of the inputs that define the test cycle followed by the average  
 665 amount of  $\text{NO}_x$  gas measured during the test. Thus, formally, we get  $\text{In} = \mathbb{R}$ ,  $\text{Out} = \mathbb{R}$ ,  
 666  $\text{Std} = \{\text{NEDC} \cdot o\}$ ,<sup>3</sup> and  $d_{\text{In}}$  and  $d_{\text{Out}}$  as defined in Example 4 [19].

667 The STL formulas developed in the previous section, combined with the probabilistic  
 668 falsification approach, give rise to further improvements to the existing testing-based work  
 669 [19] on diesel doping detection.

670 To use the falsification algorithm in Algorithm 1, we implement the restriction of the input  
 671 space to  $\text{In}_{\{\text{NEDC} \cdot o\}, \kappa_i}$  as explained in Sect. 3. With this restriction the STL formula  $\varphi_{\text{u-rob}}$   
 672 from Theorem 5 can be simplified to

$$673 \quad \square(d_{\text{Out}}((\text{NEDC} \cdot o) \downarrow_o, w \downarrow_o) - \kappa_o \leq 0). \quad (7)$$

674 This is because the conjunction and disjunction over standard traces becomes obsolete for  
 675 only a single standard trace. For the same reason, the requirement  $\square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0)$   
 676 becomes obsolete, as the compared traces are always identical. In the  $\mathcal{W}$  subformula, the  
 677 right proposition is always false, because of the restricted input space. We implemented  
 678 Algorithm 1 for the robustness computation according to formula (7).

679 <sup>3</sup> NEDC is the sequence of 1180 inputs with the  $k$ th input defining the speed of the car after  $k$  seconds from  
 680 the beginning of the NEDC

In practice, running tests like NEDC with real cars is a time consuming and expensive endeavour. Furthermore, tests on chassis dynamometers are usually prohibited to be carried out with rented cars by the rental companies. On the other hand, car emission models for simulation are not available to the public—and models provided by the manufacturer cannot be considered trustworthy. To carry out our experiments, we instead use an approximation technique that estimates the amount of  $\text{NO}_x$  emissions of a car along a certain trajectory based on data recorded during previous trips with the same car, sampled at a frequency of 1 HZ (one sample per second). Notably, these trips do not need to have much in common with the trajectory to be approximated. A trip is represented as a finite sequence  $\vartheta \in (\mathbb{R} \times \mathbb{R} \times \mathbb{R})^*$  of triples, where each such triple  $(v, a, n)$  represents the speed, the acceleration, and the (absolute) amount of  $\text{NO}_x$  emitted at a particular time instant in the sample. Speed and acceleration can be considered as the main parameters influencing the instant emission of  $\text{NO}_x$ . This is, for instance, reflected in the regulation [67, 124] where the decisive quantities to validate test routes for real-world driving emissions tests on public roads are speed and acceleration.

A recording  $\mathcal{D}$  is the union of finitely many trips  $\vartheta$ . We can turn such a recording into a predictor  $\mathcal{P}$  of the  $\text{NO}_x$  values given pairs of speed and acceleration as follows:

$$\mathcal{P}(v, a) = \text{average}[n \mid (\exists v', a'. (|v - v'| \leq 2 \wedge |a - a'| \leq 2 \wedge (v', a', n) \in \mathcal{D}))].$$

The amount of  $\text{NO}_x$  assigned to a pair  $(v, a)$  here is the average of all  $\text{NO}_x$  values seen in the recording  $\mathcal{D}$  for  $v \pm \ell$  and  $a \pm \ell$ , with  $0 \leq \ell \leq 2$ . To overcome measurement inaccuracies and to increase the robustness of the approximated emissions, the speed and acceleration may deviate up to 2 km/h, and 2 m/s<sup>2</sup>, respectively. This tolerance is adopted from the official NEDC regulation [128], which allows up to 2km/h of deviations while driving the NEDC.

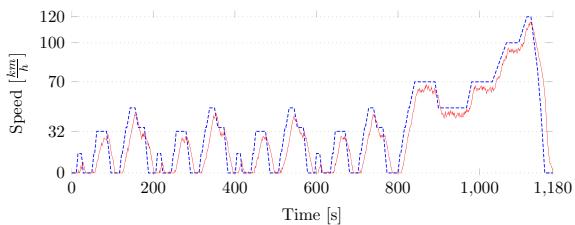
To demonstrate the practical applicability of our implementation of Algorithm 1 and our  $\text{NO}_x$  approximation, we report here on experiments with an Audi A6 Avant Diesel admitted in June 2020 as well as with its successor model admitted in 2021. We will refer to the former as car A20 and to the latter as car A21. We used the app LolaDrives to perform in total six low-cost RDE tests—two with A20 and four<sup>4</sup> with A21—and recorded the data received from the cars’ diagnosis ports. The raw data is available on Zenodo [15]. Using the emissions predictor proposed above we estimate that for an NEDC test A20 emits 86 mg/km of  $\text{NO}_x$  and that A21 emits 9 mg/km. Car A20 has previously been falsified w.r.t. the RDE specification. Neither A20 nor A21 has been falsified w.r.t. robust cleanliness.

Before turning to falsification, we spell out meaningful contexts for robust cleanliness. We identified suitable  $\text{In}$ ,  $\text{Out}$ ,  $\text{Std}$ ,  $d_{\text{In}}$ , and  $d_{\text{Out}}$  at the beginning of the section. For  $\kappa_i$ , it turned out that  $\kappa_i = 15$  km/h is a reasonable choice, as it leaves enough flexibility for human-caused driving mistakes and intended deviations [19]. The threshold for  $\text{NO}_x$  emissions under lab conditions is 80mg/km. The emission limits for RDE tests depend on the admission date of the car. Cars admitted in 2020 or earlier, must emit 168 mg/km at most, and cars admitted later must adhere to the limit of 120 mg/km. For our experiments, we use  $\kappa_o = 88$  mg/km for A20 and  $\kappa_o = 40$  mg/km for A21 to have the same tolerances as for RDE tests. Effectively, the upper threshold for A20 is  $84 + 88 = 172$  mg/km, and for A21 the limit is  $9 + 40 = 49$  mg/km. Notice that for software doping analysis, the output observed for a certain standard behaviour and the constant  $\kappa_o$  define the effective threshold; this threshold is typically different from the thresholds defined by the regulation.

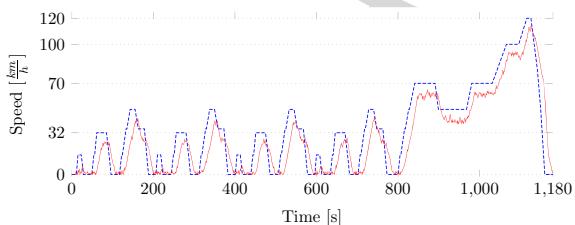
We modified Algorithm 1 by adding a timeout condition: if the algorithm is not able to find a falsifying counterexample within 3,000 iterations, it terminates and returns both the

<sup>4</sup> We do not consider test A21.3 in this article, see [22, Section 5] for details

**Fig. 2** NEDC speed profile (blue, dashed) and input falsifying  $C$  for  $\kappa_o = 88 \text{ mg/km}$  (red) with  $182 \text{ mg/km}$  of emitted  $\text{NO}_x$



**Fig. 3** NEDC speed profile (blue, dashed) and input maximising  $\text{NO}_x$  emissions to  $11 \text{ mg/km}$  (red)



trace for which the smallest robustness has been observed and its corresponding robustness value. Hence, if falsification of robust cleanliness for a system is not possible, the algorithm outputs an upper bound on how robust the system satisfies robust cleanliness.

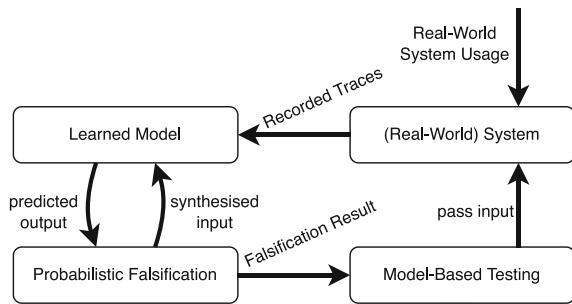
For the concrete case of the diesel emissions, the robustness value during the first 1180 inputs (sampled from the restricted input space  $\ln_{\text{Std}, \kappa_i}$ ) is always  $\kappa_o$ . When the NEDC output  $\sigma_{\text{NEDC}}$  and the non-standard output  $\sigma$  are compared, the robustness value is  $\kappa_o - |\sigma_{\text{NEDC}} - \sigma|$  (cf., eq. (7), the quantitative semantics of STL, and definition of  $d_{\text{out}}$ ). Hence, for test cycles with small robustness values, we get  $\text{NO}_x$  emissions  $\sigma$  that are either very small or very large compared to  $\sigma_{\text{NEDC}}$ . We ran the modified Algorithm 1 on A20 and A21 for the contexts defined above. For A20, it found a robustness value of  $-8$ , i.e., it was able to falsify robust cleanliness relative to the assumed contract and found a test cycle for which  $\text{NO}_x$  emissions of  $182 \text{ mg/km}$  are predicted. The test cycle is shown in Fig. 2. For A21, the smallest robustness estimate found—even after 100 independent executions of the algorithm—was  $38$ , i.e., A21 is predicted to satisfy robust cleanliness with a very high robustness estimate. The corresponding test cycle is shown in Fig. 3.

#### On Doping Tests for Cyber-physical Systems

The proposed probabilistic falsification approach to find instances of software doping needs several hundreds of iterations. This is problematic for testing real-world cyber-physical systems (CPS) to which inputs cannot be passed in an automated way. To conduct a test with a car, for example, the input to the system is a test cycle that is passed to the vehicle by driving it. Notably, we consider here the scenario that the CPS is tested by an entity that is different from the manufacturer. While the latter might have tools to overcome these technical challenges, the former typically does not have access to them.

We propose the following *integrated testing approach* for effective doping tests of cyber-physical systems. The big picture is provided in Fig. 4. In a first step, the CPS is used under real-world conditions without enforcing any specific constraints on the inputs to the system. For all executions, the inputs and outputs are recorded. So, essentially, the system can be used as it is needed by the user, but all interactions with it are recorded. From these recordings, a *model* can be learned that for arbitrary inputs (whether they were covered in the recorded data or not) predicts the output of the system. Such learning can be as simple as using statistics as we did for the emissions example above, or as complex as using deep neural nets. For the

**Fig. 4** Integrated testing approach



learned model, the probabilistic falsification algorithm computes a test input that falsifies it—inputs to this model can be passed automatically and an output is produced almost instantly. The resulting input serves as an input for the real CPS. If the prediction was correct, also the real system is falsified. If it was incorrect, the learned model can be refined and the process starts again.

For diesel emissions, the first part of this integrated testing approach has been carried out as part of the work reported in this article. We leave the second part—evaluating the generated test traces from Figs. 2 and 3 with a real car—for future work.

#### Technical Context

Software doping theory provides a formal basis for enlarging the requirements on vehicle exhaust emissions beyond too narrow lab test conditions. That conceptual limitation has by now been addressed by the official authorities responsible for car type approval [124, 127]: The old NEDC-based test procedure is replaced by the newer *Worldwide Harmonised Light Vehicles Test Procedure* (WLTP), which is deemed to be more realistic. WLTP replaces the NEDC test by a new WLTC test, but WLTC still is just a single test scenario. In addition, WLTP embraces so called *Real Driving Emissions* (RDE) tests to be conducted on public roads. A recently launched mobile phone app [20, 22], LolaDrives, harvests runtime monitoring technology for making low-cost RDE tests accessible to everyone.

Learning or approximating the behaviour of a system under test has been studied intensively. Meinke and Sindhu [82] were among the first to present a testing approach incrementally learning a Kripke structure representing a reactive system. Volpato and Tretmans [130] propose a learning approach which gradually refines an under- and over-approximation of an input-output transition system representing the system under test. The correctness of this approach needs several assumptions, e.g., an oracle indicating when, for some trace, all outputs, which extend the trace to a valid system trace, have been observed.

## 5 Individual fairness of systems evaluating humans

Example 2 introduces a new application domain for cleanliness definitions. Unica uses an AI system that is supposed to assist her with the selection of applicants for a hypothetical university. Cleanliness of such a system can be related to the fair treatment of the humans that are evaluated by it. A usable fairness analysis can happen no later than at runtime, since Unica needs to make a timely decision on whether to include the applicant in further considerations. We describe technical measures that help in mitigating this challenge by providing her with information from an individual fairness analysis in a suitable, purposeful, expedient way. To this end, we propose a formal definition for individual fairness extending the one by

[35] and based on func-cleanness. We develop a runtime monitor that analyses every output of P immediately after P's decision, which strategically searches for unfair treatment of a particular individual by comparing them to relevant hypothetical alternative individuals so as to provide a fairness assessment in a timely manner.

Much like P is to support Unica, AI systems—in the broadest sense of the word—more and more often support human decision makers. Undoubtedly, such systems should be compliant with applicable law (such as the future European AI Act [40, 41] or the Washington State facial recognition law [132]) and ought to minimise any risks to health, safety or fundamental rights. Sometimes, we cannot mitigate all these risks in advance by technical measures and also some risk-mitigation requires trade-off decisions involving features that are either impossible or difficult to operationalise and formalise. This is why it is essential that a human effectively oversees the system (which is also emphasised by several institutions such as UNESCO [129] and the European High Level Expert Group [59]). *Effective* human oversight, however, is only possible with the appropriate technical measures that allow human overseers to better understand the system at runtime [70, 71]. From a technical point of view, this raises the pressing question of what such technical measures can and ought to look like to actually enable humans to live up to these responsibilities. Our contribution is intended to bridge the gap between the normative expectations of law and society and the current reality of technological design.

## 5.1 Positioning within related research topics

Our contribution draws on and adds to three vibrant topics of current research, namely Explainable AI (XAI), AI fairness, and discrimination.

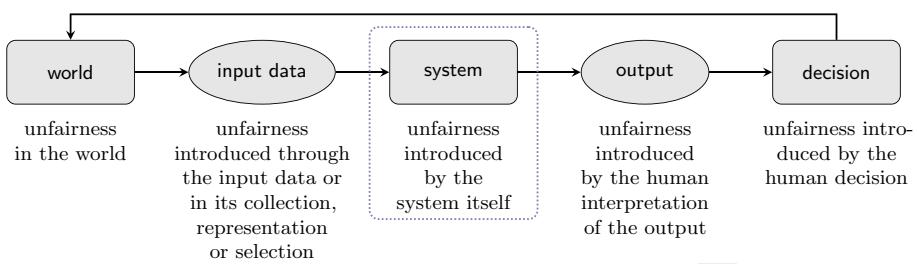
### XAI

Many of the most successful AI systems today are some kind of black boxes [11]. Accordingly, the field of ‘Explainable AI’ [54] focuses on the question of how to provide users (and possibly other stakeholders) with more information via several key perspicuity properties [117] of these systems and their outputs to make them understand these systems and their outputs in ways necessary to meet various desiderata [5, 28, 69, 74, 85, 91]. The concrete expectations and promises associated with various XAI methods are manifold. Among them are enabling warranted trust in systems [12, 62, 65, 102, 111], increasing human-system decision-making performance [68] for instance through increasing human situation awareness when operating systems [109], enabling responsible decision-making and effective human oversight [14, 80, 114], as well as identifying and reducing discrimination [74]. It often remains unclear what kind of explanations are generated by the various explainability methods and how they are meant to contribute to the fulfilment of the desiderata, even though these questions have become the subject of systematic and interdisciplinary research [69, 103].

Our approach can be taxonomised along at least two different distinctions [70, 86, 101, 102, 116]: First, it is *model-agnostic* (not *model-specific*), i.e., it is not tailored to a particular class of models but operates on observable behaviour—the inputs and outputs of the model. Second, our method is a *local method* (not *global*), i.e., it is meant to shed light on certain outputs rather than the system as a whole.

### (Un-)Fair Models

Fairness, discrimination, justice, equal opportunity, bias, prejudice, and many more such concepts are part of a meaningfully interrelated cluster that has been analysed and dissected for millennia [6, 7]. Many fields are traditionally concerned with the concepts of fairness and discrimination, ranging from philosophy [6, 7, 36, 52, 98–100] to legal sciences [25, 57, 125,



**Fig. 5** Sketch of different origins of unfairness in a decision process supported by a system; the dotted box indicates which unfairness our monitoring targets

836 [131], to psychology [60, 136], to sociology [2, 63], to political theory [99], to economics  
 837 [55]. Nowadays, it has also become a technological topic that calls for cross-disciplinary  
 838 perspectives [50].

839 With regard to fairness, there are two distinctions that are especially relevant to our work.  
 840 First, one distinction is made between *individual fairness*, i.e., that similar individuals are  
 841 treated similarly [35], and *group fairness*, i.e., that there is adequate group parity [23]. Mea-  
 842 sures of individual fairness are often close to the Aristotelian dictum to treat like cases alike  
 843 [6, 7]. In a sense, operationalisations of individual fairness are robustness measures [24, 118],  
 844 but instead of requiring robustness with respect to noise or adversarial attacks, measures of  
 845 individual fairness, such as the one by Dwork et al. [35], call for robustness with respect to  
 846 highly context-dependent differences between representations of human individuals. Second,  
 847 recent work from the field of law [131] suggests to differentiate between *bias preserving* and  
 848 *bias transforming* fairness metrics. Bias preserving fairness metrics seek to avoid adding new  
 849 bias. For such metrics, historic performances are the benchmarks for models, with equivalent  
 850 error rates for each group being a constraint. In contrast, bias transforming metrics do not  
 851 accept existing bias as a given or neutral starting point, but aim at adjustment. Therefore,  
 852 they require to make a ‘positive normative choice’ [131], i.e. to actively decide which biases  
 853 the system is allowed to exhibit, and which it must not exhibit.

854 Over the years, many concrete approaches have been suggested to foster different kinds  
 855 of fairness in artificial systems, especially in AI-based ones [74, 81, 96, 131, 134]. Yet, to  
 856 the best of our knowledge, an approach like ours is still missing. One of the approaches that  
 857 is closest to ours, namely that by John et al. [64], is not local and therefore not suitable for  
 858 runtime monitoring. Also, it is not model-agnostic. So, to the best of our knowledge, our  
 859 approach provides a new contribution to the debate on unfairness detection.

860 It is important to note/recognise that our approach can only be understood as part of a  
 861 more holistic approach to preventing or reducing unfairness. After all, there are many sources  
 862 of unfairness [8] (also see Fig. 5 and Appendix B). Therefore, not every technical measure  
 863 is able to detect every kind of unfairness and eliminating one source of unfairness might not  
 864 be sufficient to eliminate all unfairness. Our approach tackles only unfairness introduced by  
 865 the system, but not other kinds of unfairness.

### 866 **Discrimination**

867 We understand discrimination as dissimilar treatment of similar cases or similar treatment of  
 868 dissimilar cases without justifying reason. This is a definition that can also be found in the  
 869 law [44, §43]. Our work is exclusively focused on discrimination *qua* dissimilar treatment of  
 870 similar cases. Discrimination requires a thoughtful and largely not formalisable consideration  
 871 of ‘justifying reason’. However, we will exploit the relation of discrimination and fairness:

872 Unfairness in a system can arguably be a good proxy of discrimination—even though not  
 873 every unfair treatment by a system necessarily constitutes discrimination (especially not in  
 874 the legal sense). Thus, a tool that highlights cases of unfairness in a system can be highly  
 875 instrumental in detecting discriminatory features of a system. It is not viable, though, to let  
 876 such a tool rule out unfair treatment fully automatically without human oversight, since there  
 877 could be justifying reason to treat two similar inputs in a dissimilar way.

## 878 5.2 Individual fairness

879 Unica from Example 2 should be able to detect individual unfairness. An operationalisation  
 880 thereof by Dwork et al. [35] is based on the Lipschitz condition to enforce that similar  
 881 individuals are treated similarly. To measure similarity, they assume the existence of an input  
 882 distance function  $d_{\text{In}}$  and an output distance function  $d_{\text{Out}}$ . This assumption is very similar  
 883 to the one that we implicitly made in the previous sections for robust cleanliness and func-  
 884 cleanness. However, in the case of the fair treatment of humans finding reasonable distance  
 885 functions is more challenging than it was for the examples in the previous chapters. Dwork et  
 886 al. assume that both distance functions perfectly measure distances between individuals<sup>5</sup> and  
 887 between outputs of the system, respectively, but admit that in practice these distance functions  
 888 are only approximations of a ground truth at best. They suggest that distance measures might  
 889 be learned, but there is no one-size-fits-all approach to selecting distance measures. Indeed,  
 890 obtaining such distance metrics is a topic of active research [61, 87, 135]. Additionally, the  
 891 Lipschitz condition assumes a Lipschitz constant  $L$  to establish a linear constraint between  
 892 input and output distances.

893 **Definition 8** A deterministic sequential program  $P : \text{In} \rightarrow \text{Out}$  is *Lipschitz-fair* w.r.t.

894  $d_{\text{In}} : \text{In} \times \text{In} \rightarrow \mathbb{R}$ ,  $d_{\text{Out}} : \text{Out} \times \text{Out} \rightarrow \mathbb{R}$ , and a Lipschitz constant  $L$ , if and only if for  
 895 all  $i_1, i_2 \in \text{In}$ ,  $d_{\text{Out}}(P(i_1), P(i_2)) \leq L \cdot d_{\text{In}}(i_1, i_2)$ .

896 Lipschitz-fairness comes with some restrictions that limit its suitability for practical appli-  
 897 cation:

898  **$d_{\text{In}}\text{-}d_{\text{Out}}$ -relation:** High-risk systems are typically complex systems and ask for more com-  
 899 plex fairness constraints than the linearly bounded output distances  
 900 provided by the Lipschitz condition. For example, using the Lipschitz  
 901 condition prevents us from allowing small local jumps in the output and  
 902 at the same time forbidding jumps of the same rate of increase over  
 903 larger ranges of the input space (also see supplementary material in Sec-  
 904 tion [Appendix A](#)).

905 ***Input relevance:*** The condition quantifies over the entire input domain of a program. This  
 906 overlooks two things: first, it is questionable whether each input in such  
 907 a domain is plausible as a representation for a real-world individual. But  
 908 whether a system is unfair for two implausible and purely hypothetical  
 909 inputs is largely irrelevant in practice. Secondly, it also ignores that mere  
 910 potential unfair treatment is at most a threat, not necessarily already a  
 911 harm [106]. Therefore, even with a restriction to only plausible appli-  
 912 cants, the analysis might take into account more inputs than needed for

<sup>5</sup> For easier readability, we will not distinguish between *individuals* and their *representations* unless this distinction is relevant in the specific context. It is nevertheless important to note that inputs are not individuals, but only representations of individuals, since an input could inadequately represent an individual and therefore be unfair (also see [Appendix B](#)).

many real-world applications. What is important in practice is the ability to determine whether *actual* applicants are treated unfairly—and for this it is often not needed to look at the entire input domain.

**Monitorability:** In a monitoring scenario with the Lipschitz condition in place, a fixed input  $i_1$  must be compared to potentially all other inputs  $i_2$ . Since the input domain of the system can be arbitrarily large, the Lipschitz condition is not yet suitable for monitoring in practice (for a related point see John et al. [64]).

We propose a notion of individual fairness that is based on Definition 3. Instead of cleanliness contracts we consider here *fairness contracts*, which are tuples  $\mathcal{F} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$  containing input and output distance functions and the function  $f$  relating input distances and output distances. Notably, the set of standard inputs  $\text{StdIn}$  known from cleanliness contracts is not part of a fairness contract; it is unknown what qualifies an input to be ‘standard’ in the context of fairness analyses. Still, our fairness definition evaluates fairness for a set of individuals  $\mathcal{I} \subseteq \text{In}$  (e.g., a set of applicants), which has conceptual similarities to the set  $\text{StdIn}$ . A fairness contract specifies certain fairness parameters for a concrete context or situation. Such parameters should generally not already include  $\mathcal{I}$  to avoid introducing new unfairness through the monitor by tailoring it to specific inputs individually or by treating certain inputs differently from others. Func-fairness can thus be defined as follows:

**Definition 9** A deterministic sequential program  $P : \text{In} \rightarrow \text{Out}$  is *func-fair* for a set  $\mathcal{I} \subseteq \text{In}$  of actual inputs w.r.t. a fairness contract  $\mathcal{F} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$ , if and only if for every  $i \in \mathcal{I}$  and  $i' \in \text{In}$ ,  $d_{\text{Out}}(P(i), P(i')) \leq f(d_{\text{In}}(i, i'))$ .

The idea behind func-fairness is that every individual in set  $\mathcal{I}$  is compared to potential other inputs in the domain of  $P$ . These other inputs do not necessarily need to be in  $\mathcal{I}$ , nor do these inputs need to have “physical counterparts” in the real world. Driven by the insights of the *Input relevance* restriction of Lipschitz-fairness, we explicitly distinguish inputs in the following and will call inputs that are given to  $P$  by a user *actual inputs*, denoted  $i_a$ , and call inputs to which such  $i_a$  are compared to *synthetic inputs*, denoted  $i_s$ . Actual inputs are typically<sup>6</sup> inputs that have a real-world counterpart, while this might or might not be true for synthetic inputs. On first glance, an alternative to using synthetic inputs is to use only actual inputs, e.g., to compare every actual input with every other actual input in  $\mathcal{I}$ . For example, for a university admission, all applicants could be compared to every other applicant. However, this would heavily rely on contingencies: the detection of unfair treatment of an applicant depends on whether they were lucky enough that, coincidentally, another candidate has also applied who aids in unveiling the system’s unfairness towards them. Instead, func-fairness prefers to over-approximate the set of plausible inputs that actual inputs are compared to rather than under-approximating it by comparing only to other inputs in  $\mathcal{I}$ . This way, the attention of the human exercising oversight of the system might be drawn to cases that are actually not unfair, but as a competent human in the loop, they will most likely be able to judge that the input was compared to an implausible counterpart. This will usually enable more effective human oversight than an under-approximation that misses to alert the human to unfair cases.

Notice that func-fairness is a conservative extension of Lipschitz-fairness. With  $\mathcal{I} = \text{In}$  and  $f(x) = L \cdot x$ , func-fairness mimics Lipschitz-fairness. Wachter et al. [131] classify the Lipschitz-fairness of Dwork et al. [35] as bias-transforming. As we generalise this and

<sup>6</sup> A case where actual inputs might not have real-world counterparts is testing.

---

**Algorithm 2** FairnessMonitor, with  $\xi$ -min  $S = (\xi, i_1, i_2)$  only if  $(\xi, i_1, i_2) \in S$  and for all  $(\xi', i'_1, i'_2) \in S$ ,  $\xi' \geq \xi$

---

**Falsification Parameters:** PS: Proposal scheme,  $\beta$ : Temperature parameter

**Input:** System  $P : In \rightarrow Out$ , Fairness contract  $\mathcal{F} = \langle d_{In}, d_{Out}, f \rangle$ , and set of actual inputs  $\mathcal{I}$

**Output:** A minimal fairness score triple from  $\mathbb{R} \times \mathcal{I} \times In$ .

```

1:  $i_s \leftarrow$  any input  $i_a \in \mathcal{I}$ 
2:  $(\xi, i_{\min}, i_s) \leftarrow \xi\text{-min}\{(F(i_a, i_s), i_a, i_s) \mid i_a \in \mathcal{I}\}$ 
3:  $(\xi_{\min}, i_1, i_2) \leftarrow (\xi, i_{\min}, i_s)$ 
4: while not timeout do
5:    $i'_s \leftarrow PS(i_s, P(i_s))$ 
6:    $(\xi', i'_{\min}, i'_s) \leftarrow \xi\text{-min}\{(F(i_a, i'_s), i_a, i'_s) \mid i_a \in \mathcal{I}\}$ 
7:    $(\xi_{\min}, i_1, i_2) \leftarrow \xi\text{-min}\{(\xi_{\min}, i_1, i_2), (\xi', i'_{\min}, i'_s)\}$ 
8:    $\alpha \leftarrow \exp(-\beta(\xi' - \xi))$ 
9:    $r \leftarrow \text{UniformRandomReal}(0, 1)$ 
10:  if  $r \leq \alpha$  then
11:     $i_s \leftarrow i'_s$ 
12:     $\xi \leftarrow \xi'$ 
13:  end if
14: end while
15: return  $(\xi_{\min}, i_1, i_2)$ 

```

---

958 introduce no element that has to be regarded as bias-preserving, our approach arguably is  
 959 bias-transforming, too.

960 Func-fairness, with its function  $f$ , provides a powerful tool to model complex fairness  
 961 constraints. How such an  $f$  is defined has profound impact on the quality of the fairness  
 962 analysis. A full discussion about which types of functions make a good  $f$  go beyond the  
 963 scope of this article. A suitable choice for  $f$  and the distance functions  $d_{In}$  and  $d_{Out}$  heavily  
 964 depends on the context in which fairness is analysed—there is no one-fits-it-all solution.  
 965 Func-fairness makes this explicit with the formal fairness contract  $\mathcal{F} = \langle d_{In}, d_{Out}, f \rangle$ .

### 966 5.3 Fairness monitoring

967 We develop a probabilistic-falsification-based fairness monitor that, given a set of actual  
 968 inputs, searches for a synthetic counterexample to falsify a system  $P$  w.r.t. a fairness contract  
 969  $\mathcal{F}$ . To this end, it is necessary to provide a quantitative description of func-fairness that satisfies  
 970 the characteristics of a robustness estimate. We call this description *fairness score*. For an  
 971 actual input  $i_a$  and a synthetic input  $i_s$  we define the fairness score as  $F(i_a, i_s) := f(d_{In}(i_a, i_s)) -$   
 972  $d_{Out}(P(i_a), P(i_s))$ .  $F$  is indeed a robustness estimate function: if  $F(i_a, i_s)$  is non-negative, then  
 973  $d_{Out}(P(i_a), P(i_s)) \leq f(d_{In}(i_a, i_s))$ , and if it is negative, then  $d_{Out}(P(i_a), P(i_s)) \not\leq f(d_{In}(i_a, i_s))$ .  
 974 For a set of actual inputs  $\mathcal{I}$ , the definition generalises to  $F(\mathcal{I}, i_s) := \min\{F(i_a, i_s) \mid i_a \in \mathcal{I}\}$ ,  
 975 i.e., the overall fairness score is the minimum of the concrete fairness scores of the inputs in  
 976  $\mathcal{I}$ . Notice that  $\mathcal{R}_{\mathcal{I}}(i_s) := F(\mathcal{I}, i_s)$  is essentially the quantitative interpretation of  $\varphi_{u\text{-func}}$  (from  
 977 Theorem 6) after simplifications attributed to the fact that  $P$  is a sequential and deterministic  
 978 program (cf. Definition 2.2 vs. Definition 3).

979 Algorithm 2 shows FairnessMonitor, which builds on Algorithm 1 to search for the mini-  
 980 mal fairness score in a system  $P$  for fairness contract  $\mathcal{F}$ . The algorithm stores fairness scores  
 981 in triples that also contain the two inputs for which the fairness score was computed. The  
 982 minimum in a set of such triples is defined by the function  $\xi$ -min that returns the triple with  
 983 the smallest fairness score of all triples in the set. The first line of FairnessMonitor initialises  
 984 the variable  $i_s$  with an arbitrary actual input from  $\mathcal{I}$ . For this value of  $i_s$ , the algorithm checks

**Algorithm 3** FairnessAwareSystem

**Parameters:** System  $P : \text{In} \rightarrow \text{Out}$ , Fairness contract  $\mathcal{F} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$

**Input:** Input  $i_a \in \text{In}$

**Output:** Tuple of the system output, normalised fairness score, and synthetic values witnessing the fairness score

```
1:  $(\xi_{\min}, i_a, i_s) \leftarrow \text{FairnessMonitor}(P, \mathcal{F}, \{i_a\})$ 
2: return( $P(i_a), \xi_{\min} \div f(d_{\text{In}}(i_a, i_s)), (i_s, P(i_s))$ )
```

the corresponding fairness scores for all actual inputs  $i_a \in \mathcal{I}$  and stores the smallest one. In line 3, the globally smallest fairness score triple is initialised. In line 5 it uses the proposal scheme to get the next synthetic input  $i'_s$ . Line 6 is similar to line 2: for the newly proposed  $i'_s$  it finds the smallest fairness score, stores it, and updates the global minimum if it found a smaller fairness score (line 7). Lines 8-13 come from Algorithm 1. The only difference is that in addition to  $i_s$  we also store the fairness score  $\xi$ . Line 4 of Algorithm 2 differs from Algorithm 1 by terminating the falsification process after a timeout occurs (similar to the adaptation of Algorithm 1 in Sect. 4). Hence, the algorithm does not (exclusively) aim to falsify the fairness property, but aims at minimising the fairness score; even if the fair treatment of the inputs in  $\mathcal{I}$  cannot be falsified in a reasonable amount of time, we still learn how robustly they are treated fairly, i.e., how far the least fairly treated individual in  $\mathcal{I}$  is away from being treated unfairly. After the timeout occurs, the algorithm returns the triple with the overall smallest seen fairness score  $\xi_{\min}$ , together with the actual input  $i_1$  and the synthetic input  $i_2$  for which  $\xi_{\min}$  was found. In case  $\xi_{\min}$  is negative,  $i_2$  is a counterexample for  $P$  being func-fair.

FairnessMonitor implements a sound  $\mathcal{F}$ -unfairness detection as stated in Proposition 7. However, it is not complete, i.e., it is not generally the case that  $P$  is func-fair for  $\mathcal{I}$  if  $\xi$  is positive. It may happen that there is a counterexample, but FairnessMonitor did not succeed in finding it before the timeout. This is analogue to results obtained for model-agnostic robust cleanliness analysis [19].

**Proposition 7** Let  $P : \text{In} \rightarrow \text{Out}$  be a deterministic sequential program,  $\mathcal{F} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$  a fairness contract, and  $\mathcal{I}$  a set of actual inputs. Further, let  $(\xi_{\min}, i_1, i_2)$  be the result of FairnessMonitor( $P, \mathcal{F}, \mathcal{I}$ ). If  $\xi_{\min}$  is negative, then  $P$  is not func-fair for  $\mathcal{I}$  w.r.t.  $\mathcal{F}$ .

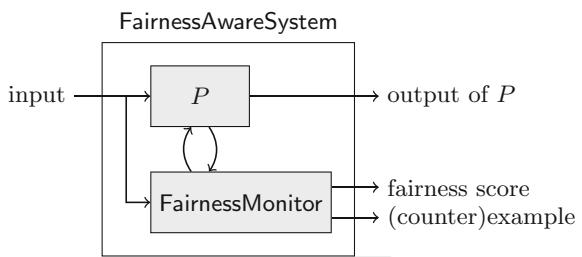
Moreover, FairnessMonitor circumvents major restrictions of the Lipschitz-fairness:

**1009     $d_{\text{In}}\text{-}d_{\text{Out}}$ -relation:** Func-fairness defines constraints between input and output distances by  
1010    means of a function  $f$ , which allows to express also complex fairness  
1011    constraints. For a more elaborate discussion, see Sect. Appendix A.

**1012    Input relevance:** Func-fairness explicitly distinguishes between actual and synthetic  
1013    inputs. This way, func-fairness acknowledges a possible obstacle of the  
1014    fairness theory when it comes to a real-world usage of the analysis,  
1015    namely that only some elements of the system's input domain might be  
1016    plausible and that usually only few of them become actual inputs that  
1017    have to be monitored for unfairness.

**1018    Monitorability:** FairnessMonitor demonstrates that func-fairness is monitorable. It  
1019    resolves the quantification over  $\text{In}$  using the above concepts from proba-  
1020    bilistic falsification using the robustness estimate function  $F$  as defined  
1021    above.

**Fig. 6** Schematic visualisation of FairnessAwareSystem



### 1022 **Towards func-fairness in the loop**

1023 If a high-risk system is in operation, a human in the loop must oversee the correct and  
 1024 fair functioning of the outputs of the system. To do this, the human needs real-time fairness  
 1025 information. Figure 6 shows how this can be achieved by coupling the system  $P$  and the  
 1026 FairnessMonitor in Algorithm 2 in a new system called FairnessAwareSystem.  
 1027 FairnessAwareSystem is sketched in Algorithm 3. Intuitively, the FairnessAwareSystem is  
 1028 a higher-order program that is parameterised with the original program  $P$  and the fairness con-  
 1029 tract  $\mathcal{F}$ . When instantiated with these parameters, the program takes arbitrary (actual) inputs  
 1030  $i_a$  from  $\text{In}$ . In the first step, it does a fairness analysis using FairnessMonitor with arguments  
 1031  $P$ ,  $\mathcal{F}$ , and  $\{i_a\}$ . To make fairness scores comparable, FairnessAwareSystem normalises the  
 1032 fairness score  $\xi$  received from FairnessMonitor by dividing<sup>7</sup> it by the output distance limit  
 1033  $f(d_{\text{In}}(i_a, i_s))$ . For fair outputs, the score will be between 0 (almost unfair) and 1 (as fair as  
 1034 possible).<sup>8</sup> Outputs that are not func-fair are accompanied by a negative score representing  
 1035 how much the limit  $f(d_{\text{In}}(i_a, i_s))$  is exceeded. A fairness score of  $-n$  means that the output  
 1036 distance of  $P(i_a)$  and  $P(i_s)$  is  $n + 1$  times as high as that limit. Finally, FairnessAwareSystem  
 1037 returns the triple with  $P$ 's output for  $i_a$ , the normalised fairness score, and the synthetic input  
 1038 with its output witnessing the fairness score.

### 1039 **Interpretation of monitoring results**

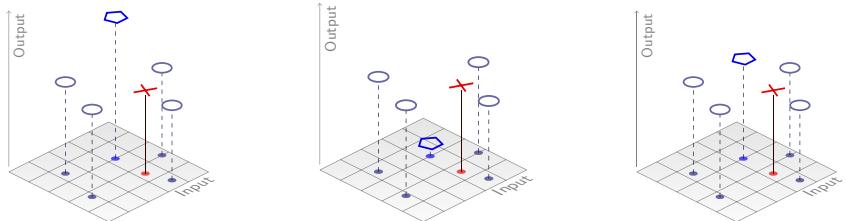
1040 Especially when FairnessAwareSystem finds a violation of func-fairness, the suitable inter-  
 1041 pretation and appropriate response to the normalised fairness score proves to be a non-trivial  
 1042 matter that requires expertise.

1043 **Example 6** Instead of using  $P$  from Example 2 on its own, Unica now uses FairnessAwareSystem  
 1044 with a suitable fairness contract. and thereby receive a fairness score along with  $P$ 's verdict  
 1045 on each applicant. (Which fairness contracts are suitable is an open research problem, see  
 1046 *Limitations & Challenges* in Sect. 7.) If the fairness score is negative, she can also take into  
 1047 account the information on the synthetic counterpart returned by FairnessAwareSystem.  
 1048 Among the 4096 applicants for the PhD program, the monitoring assigns a negative fairness  
 1049 score to three candidates: Alexa, who received a low score, Eugene, who was scored very  
 1050 highly, and John, who got an average score. According to their scoring, Alexa would be  
 1051 desk-rejected, while Eugene and John would be considered further.

1052 Alexa's synthetic counterpart, let's call him Syntbad, is ranked much higher than Alexa.  
 1053 In fact, he is ranked so high that Syntbad would not be desk-rejected. Unica compares Alexa

<sup>7</sup> For  $f$  that can return 0, there may be a  $0 \div 0$  division. The result of this division should be defined depending on the concrete context; reasonable values range from the extreme scores 0 (to indicate that the score is on the edge to becoming 'unfair') to 1 (to indicate that more fairness is impossible).

<sup>8</sup> Fairness may be a vague concept that cannot be dichotomised. By its choice of the fairness contract parameters, our approach nevertheless specifies a (non-arbitrary) cut-off point at 0; but it does so for purely instrumental and non-ontological reasons.



(a) case of unfairness where input is treated worse than relevant counterpart

(b) case of unfairness where input is treated better than relevant counterpart

(c) case of no detected unfairness

**Fig. 7** Exemplary illustration of configurations of an input (red cross) and its synthetic counterparts (grey circles) and the synthetic counterpart with the minimal fairness score (blue polygon); with a two-dimensional input space (grid) and a one-dimensional output

and Synbad and finds that they only differ in one respect: Synbad's graduate university is the one in the official ranking that is immediately *below* the one that Alexa attended. Unica does some research and finds that Alexa's institution is predominantly attended by People of Colour, while this is not the case for Synbad's institution. Therefore, FairnessAwareSystem helped Unica not only to find an unfair treatment of Alexa, but also to uncover a case of potential racial discrimination.

John's counterpart, Sinclair, is ranked much lower than him. Unica manually inspects John's previous institution (an infamous online university), his GPA of 1.8, and his test result with only 13%. She finds that this very much suggests that John will not be a successful PhD candidate and desk-rejects him. Therefore, Unica has successfully used FairnessAwareSystem to detect a fault in scoring system P whereby John would have been treated unfairly in a way that would have been to his advantage.

Eugene received a top score, but his synthetic counterpart, Syna, received only an average one. Unica suspects that Eugene was ranked too highly given his graduate institution, GPA, and test score. However, as he would not have been desk-rejected either way, nothing changes for Eugene, and the unfairness he was subject to, is not of effect to him.

The cases of John and Eugene share similarities with the configuration in (b) in Fig. 7, the one of Alexa with (a), and the ones of all other 4093 candidates with (c).

If our monitor finds only a few problematic cases in a (sufficiently large and diverse) set of inputs, our monitoring helps Unica from our running example by drawing her attention to cases that require special attention. Thereby, individuals who are judged by the system have a better chance of being treated fairly, since even rare instances of unfair treatment are detected. If, on the other hand, the number of problematic cases found is large, or Unica finds especially concerning cases or patterns, this can point to larger issues within the system. In these cases, Unica should take appropriate steps and make sure that the system is no longer used until clarity is established why so many violations or concerning patterns are found. If the system is found to be systematically unfair, it should arguably be removed from the decision process. A possible conclusion could also be that the system is unsuitable for certain use cases, e.g., for the use on individuals from a particular group. Accordingly, it might not have to be removed altogether but only needs to be restricted such that problematic use cases are avoided. In any case, significant findings should also be fed back to developers or deployers of the potentially problematic system. A fairness monitoring such as in FairnessAwareSystem or a fairness

analysis as in FairnessMonitor could also be useful to developers, regulating authorities, watchdog organisations, or forensic analysts as it helps them to check the individual fairness of a system in a controlled environment.

## 6 Interdisciplinary assessment of fairness monitoring

Regulations for car related emissions are in force for a considerable amount of time, thus, its legal interpretation is mostly clear. In case of human oversight of AI systems, the AI act is new and parts of it are legally ambiguous. This raises the question of whether our approach meets requirements that go beyond pre-theoretical deliberations. Even though comprehensive analyses would go far beyond the scope of this paper, we will nevertheless assess some key normative aspects in philosophical and legal terms, and also briefly turn to the related empirical aspects, especially from psychology.

### 6.1 Psychological assessment

Fairness monitoring promises various advantages in terms of human-system interaction in application contexts—provided it is extended by an adequate user interface—which call for empirical tests and studies. We will only discuss a possible benefit that closely aligns with the current draft of the AI Act: our approach may support effective human oversight. Two central aspects of effective oversight are situation awareness and warranted trust. Our method highlights unfairness in outputs which can be expected to increase users' situation awareness (i.e., 'the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future' [37, p. 36]), which is a variable central for effective oversight [38]. In the minimal case, this allows users to realise that something requires their attention and that they should check the outputs for plausibility and adequacy. In the optimal case and after some experience with the monitor, it may even allow users to predict instances where a system will produce potentially unfair outputs. In any case, the monitoring should enable them to understand limitations of the system and to feed back their findings to developers who can improve the system. This leads us to warranted trust, which includes that users are able to adequately judge when to rely on system outputs and when to reject them [62, 73]. Building warranted trust strongly depends on users being able to assess system trustworthiness in the given context of use [73, 110]. According to their theoretical model on trust in automation, Lee and See [73] propose that trustworthiness relates to different facets of which performance (e.g., whether the system performs reliably with high accuracy) and process (e.g., knowing how the system operates and whether the system's decision-processes help to fulfil the trustor's goals) are especially relevant in our case. Specifically, fairness monitoring should enable users to more accurately judge system performance (e.g., by revealing possible issues with system outputs) and system processes (e.g., whether the system's decision logic was appropriate). In line with Lee and See's propositions, this should provide a foundation for users to be better able to judge system trustworthiness and should thus be a promising means to promote warranted trust. In consequence, our monitoring provides a needed addition to high-risk use contexts of AI because it offers information enabling humans to more adequately use AI-based systems in the sense of possibly better human-system decision performance and with respect to user duties as described in the AI Act.

## 1128 6.2 Philosophical assessment

1129 More effective oversight promises more informed decision-making. This, in turn, enables  
 1130 morally better decisions and outcomes, since humans can morally ameliorate outcomes in  
 1131 terms of fairness and can see to it that moral values are promoted. Also, fairness monitoring  
 1132 helps in safeguarding fundamental democratic values if it is applied to potentially unfair  
 1133 systems which are used in certain societal institutions of a high-risk character such as courts  
 1134 or parliaments. It could, for example, make AI-aided court decisions more transparent and  
 1135 promote equality before the law. However, since our approach requires finding context-  
 1136 appropriate and morally permissible parameters for  $\mathcal{F}$ , moral requirements arise to enable  
 1137 the finding of such parameters. This not only affects, e.g., developers of such systems, but  
 1138 also those who are in a position to enforce that adequate parameters are chosen, such as  
 1139 governmental authorities, supervising institutions or certifiers.

1140 Apart from that, various parties have arguably a legitimate interest in adequately ascribing  
 1141 moral responsibility for the outcomes of certain decisions to human deciders [14]—regardless  
 1142 of whether the decision making process is supported by a system. Adequately ascribing moral  
 1143 responsibility is not always possible, though. One precondition for moral responsibility is  
 1144 that the agent had sufficient epistemic access to the consequences of their doing [90, 119], i.e.,  
 1145 that they have enough and sufficiently well justified beliefs about the results of their decision.  
 1146 Someone overseeing a university selection process (like Unica) should, for example, have  
 1147 sufficiently well justified beliefs that, at the very least, their decisions do not result in more  
 1148 unfairness in the world. If the admission process is supported by a black-box system, though,  
 1149 Unica cannot be expected to have any such beliefs since she lacks insight in the fairness of the  
 1150 system. Therefore, adequate responsibility ascription is usually not possible in this scenario.  
 1151 Our monitoring alleviates this problem by providing the decider with better epistemic access  
 1152 to the fairness of the system.

1153 FairnessAwareSystem helps in making Unica's role in the decision process significant  
 1154 and not only that of a mere button-pusher. FairnessAwareSystem makes it possible for her to  
 1155 fulfil some of the responsibilities and duties plausibly associated with her role. For example,  
 1156 she can now be realistically expected to not only detect, but resolve at least some cases of  
 1157 apparent unfairness competently (although she may need additional information to do so).  
 1158 In this respect, she should not be 'automated away' (cf. [79]).

## 1159 6.3 Legal assessment

1160 A central legislative debate of our time is how to counter the risks AI systems can pose to the  
 1161 health and safety or fundamental rights of natural persons. Protective measures must be taken  
 1162 at various levels: First, before being permitted on the market, it must be ensured *ex ante* that  
 1163 such high-risk AI-systems are in conformity with mandatory requirements<sup>9</sup> regarding safety  
 1164 and human rights. This means in particular that the selection of the properties which a system  
 1165 should exhibit requires a positive normative choice and should not simply replicate biases  
 1166 present in the status quo [131]. In addition, AI-systems must be designed and developed in  
 1167 such a way that natural persons can oversee their functioning. For this purpose, it is necessary  
 1168 for the provider to identify appropriate human oversight measures before its placing on the  
 1169 market or putting into service. In particular, such measures should guarantee that the natural

<sup>9</sup> The specific risks set by AI-systems may also give reason to consider an adaptation and expansion of European legal frameworks such that an even broader prohibition of discrimination (cf. "Appendix C.1") is set into place.

1170 persons to whom human oversight has been assigned have the necessary competence, training  
 1171 and authority to carry out that role [40, recital 48] [41, Art. 14 (5)].

1172 Second, during runtime, the proper functioning of high-risk AI systems, which have been  
 1173 placed on the market lawfully, must be ensured. To achieve this goal, a bundle of different  
 1174 measures is needed, ranging from legal obligations to implement and perform meaningful  
 1175 oversight mechanisms to user training and awareness in order to counteract ‘automation bias’.  
 1176 Furthermore, the AI Act proposal requires deployers to inform the provider or distributor  
 1177 and suspend the use of the system when they have identified any serious incidents or any  
 1178 malfunctioning [40, 41, Art. 29(4)].

1179 Third, and *ex post*, providers must act and take the necessary corrective actions as soon  
 1180 as they become aware, e.g. through information provided by the deployer, that the high-risk  
 1181 system does not (or no longer) meet the legal requirements [40, 41, Art. 16(g)]. To this end,  
 1182 they must establish and document a system of monitoring that is proportionate to the type of  
 1183 AI technology and the risks of the high-risk AI system [40, 41, Art. 61(1)].

1184 Fairness monitoring can be helpful in all three of the above respects. Therefore, we argue  
 1185 that there is even a legal obligation to use technical measures such as the method presented  
 1186 in this paper if this is the only way to ensure effective human oversight.

## 1187 7 Conclusion and future work

1188 This article brings together software doping theory and probabilistic falsification techniques.  
 1189 To this end, it proposes a suitable HyperSTL semantics and characterises robust cleanliness and  
 1190 func-cleanliness as HyperSTL formulas and, for the special case of finite standard behaviour,  
 1191 STL formulas. Software doping techniques have been extensively applied to the tampered  
 1192 diesel emission cleaning systems; this article continues this path of research by demonstrating  
 1193 how testing of real cars can become more effective. For the first time, we apply software  
 1194 doping techniques to high-risk (AI) systems. We propose a runtime fairness monitor to  
 1195 promote effective human oversight of high-risk systems. The development of this monitor is  
 1196 complemented by an interdisciplinary evaluation from a psychological, philosophical, and  
 1197 legal perspective.

### 1198 Limitations & Challenges

1199 A challenge to those employing robust cleanliness or func-cleanliness analysis is the selection  
 1200 of suitable parameters, especially  $d_{in}$ ,  $d_{out}$ , and  $f$  or  $\kappa_i$  and  $\kappa_o$ . Because of their high degree  
 1201 of context sensitivity, there are no paradigmatic candidates for them that one can default to.  
 1202 Instead, they have to be carefully selected with the concrete system, the structure of input  
 1203 data and the situation of use in mind.

1204 Reasonable choices for robust cleanliness analysis of diesel emissions have been proposed  
 1205 in recent work [19, 21]. With respect to individual fairness analysis, potential systems to  
 1206 which FairnessAwareSystem or FairnessMonitor can be applied to are still too diverse to  
 1207 give recommendations for the contract parameters. Obviously, further technical limitations  
 1208 include that  $f$ ,  $d_{in}$ , and  $d_{out}$  must be computable.

1209 With a particular regard to fairness analysis, we identify also non-technical limitations. As  
 1210 seen in Fig. 5, our fairness monitoring aims to uncover a particular kind of unfairness, namely  
 1211 individual unfairness that originates from within the system. This excludes all kinds of group  
 1212 unfairness as well as unfairness from sources other than the system. Another limitation is  
 1213 the human’s competence to interpret the system outputs. Even though this is not a limitation  
 1214 that is inherent to our approach, it nevertheless will arguably be relevant in some practical

cases, and an implementation of the monitoring always has to happen with the human in mind. For example, the design of the tool should avoid creating the false impression that the system is proven to be fair for an individual if no counterexample has been found. Interpretations like this could lead to inflated judgements of system trustworthiness and eventually to overtrusting system outputs [110, 112]. Also, it might be reasonable to limit access to the monitoring results: if individuals who are processed by the system have full access to their fairness analysis, they could use this to ‘game’ the system, i.e. they could use the synthetic inputs to slightly modify their own input such that they receive a better outcome. While more transparency for the user is generally desirable, this has to be kept in mind to avoid introducing new unfairness on a meta-level.

## Future Work

The probabilistic falsification technique we use in this article can be seen as a modular framework that consists of several interchangeable components. One of these components is the optimisation technique used to find the input with minimal robustness value. Algorithm 1 uses a simulated annealing technique [29, 107], but other techniques have been proposed for temporal logic falsification, too [4, 108]. We want to further look into such alternative optimisation techniques and to evaluate if they offer benefits w.r.t. cleanliness falsification.

Finally, the fairness monitoring approach has been presented using a toy example. It is not claimed to be readily applicable to real-life scenarios. Besides the future work that has already been mentioned throughout the paper, we are planning on various extensions of our approach, and are working on an implementation that will allow us to integrate the monitoring into a real system. Moreover, we plan to test the possible benefits and shortcomings of the approach in user studies where decision-makers are tasked to make hiring decisions with and without the fairness monitoring approach. Further work will encompass activities such as the improvement and embedding of the algorithm FairnessAwareSystem into a proper tool that can be used by non-computer-scientists, and the extension of the monitoring technique to cover more types of unfairness. For example, logging the output of the fairness monitor could be used to identify groups that are especially likely to be treated unfairly by the system: The individual fairness verdicts provided by FairnessAwareSystem and FairnessMonitor may also be logged and considered for further fairness assessments or other means of quality assurance of system  $P$ . Statistical analysis might unveil that individuals of certain groups are treated unfairly more frequently than individuals from other groups. Depending on the distinguishing features of the evaluated group, this can uncover problems in  $P$ , especially if protected attributes, such as gender, race, age, etc, are taken into account. Thereby, system fairness can be assessed for protected attributes without including them in the input of  $P$ , which should generally be avoided, and even without disclosing them to the human in the loop. By evaluating the monitoring logs from sufficiently many diverse runs of FairnessAwareSystem, our local method can be lifted such that it resembles a global method for many practical applications, i.e. we can make statistical statements about the general fairness of  $P$ . Such an evaluation can also be used to extract prototypes and counterexamples in the spirit of Been et al. [66] illustrating the *tendency* to judge unfairly. This is an interesting combination of individual and group fairness that we want to look into further. Other insights from the research on reactive systems [19, 21, 32] can potentially be used to further enrich the monitoring. Finally, various disciplines have to join forces to resolve highly interdisciplinary questions such as what constitutes reasonable and adequate choices for  $f$ ,  $d_{in}$ , and  $d_{out}$  in given contexts of application.

**Author Contributions** Not applicable

1262 **Funding** This work is partially funded by DFG grant 389792660 as part of TRR 248—CPEC (see <https://perspicuous-computing.science>), by VolkswagenStiftung as part of grants AZ 98514, 98513 and 98512 EIS—  
 1263 Explainable Intelligent Systems (see <https://explainable-intelligent.systems>), and by the European Regional  
 1264 Development Fund (ERDF) and the Saarland as part of CERTAIN—Center for European Research in Trusted  
 1265 AI (see <https://www.certain-trust.eu/>). It has received support as part of STORM\_SAFE, an Interreg project  
 1266 supported by the North Sea Programme of the European Regional Development Fund of the European Union.  
 1267

1268 **Data Availability** The datasets analysed during the current study are available in a Zenodo repository [15]  
 1269 (<https://zenodo.org/record/8058770>).

## 1270 Declarations

1271

1272 **Competing interests** The authors have no competing interests to declare that are relevant to the content of  
 1273 this article.

1274 **Ethics approval** Not applicable

1275 **Consent to participate** Not applicable

1276 **Consent for publication** Not applicable

1277 **Code availability** Not applicable

## 1278 Appendix A Technical Appendix

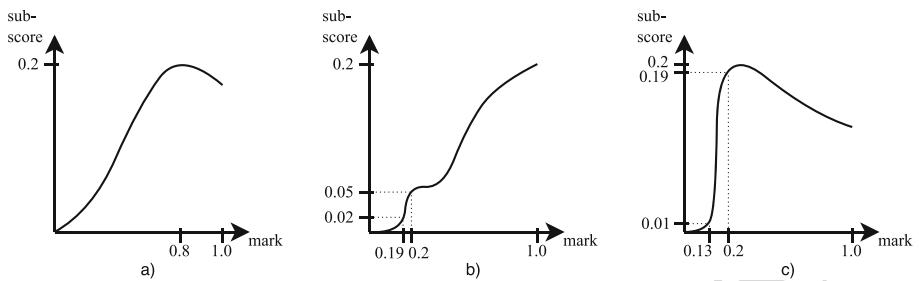
1279 This appendix illustrates that func-fairness is more expressive than Lipschitz-fairness and why  
 1280 this is useful. For this, we use as a toy example a very simple, hypothetical HR scoring system  
 1281 that aggregates five scores given to the candidates. We remark that the whole scenario, the  
 1282 implementation of the system, the choice of distance functions and  $f$ , is likely not applicable  
 1283 for real-life situations; everything is picked so that our explanations are understandable.

1284 Suppose that certain qualities and characteristics of the applicants are pre-scored by other  
 1285 systems on a scale from 0 to 100 %, where 0 means that the candidate is utterly unsuitable  
 1286 for the job in a certain regard, while a scoring of 100 % means that the candidate is perfect  
 1287 for the job in this regard. In particular, we will assume that the following marks are given to  
 1288 each applicant: an *education mark* for how well they are academically suitable for the job, an *experience mark*  
 1289 for how well their previous work experience fits the job, a *personality mark* for their personal and social skills,  
 1290 a *mental ability mark* for what is colloquially referred to as an applicant’s general intelligence, and, finally, a *skill mark* that tracks the special skills  
 1291 that applicants have which might be beneficial for the job, such as their knowledge of foreign  
 1292 languages.

1293 The system  $P$  that is of interest for us in this example is the one that aggregates all of  
 1294 these marks and gives out an overall score of how well the candidate is suited for the job.  
 1295 The human responsible for the hiring process can use this in her hiring decision, e.g., she can  
 1296 focus on the top-scoring candidates and choose among them.

1297 Let  $\mathcal{M} = [0, 1] \subseteq \mathbb{R}$  be the reals between 0 and 1. Each of the five marks mentioned above  
 1298 is a real number from set  $\mathcal{M}$ . The input domain  $\ln = \mathcal{M}^5$  for the sketched HR system is a  
 1299 tuple of five marks. The output of the system is the overall suitability score of an applicant,  
 1300 which is also a value from  $\mathcal{M}$ . The distance between two inputs is defined as the euclidean  
 1301 distance, normalised to a value between 0 and 1, i.e.,

$$1303 d_{\ln}((ed_1, ex_1, pe_1, in_1, sk_1), (ed_2, ex_2, pe_2, in_2, sk_2)) =$$



**Fig. 8** Visualisation of subscore functions mapping marks to subscores

$$1304 \quad \sqrt{\frac{(ed_1 - ed_2)^2 + (ex_1 - ex_2)^2 + (pe_1 - pe_2)^2 + (in_1 - in_2)^2 + (sk_1 - sk_2)^2}{5}},$$

1305 where  $ed$  represents the education mark,  $ex$  the experience mark,  $pe$  the personality mark,  
 1306 in the mental ability mark, and  $sk$  the skill mark of an applicant. The distance between two  
 1307 outputs  $d_{\text{out}}(o_1, o_2) = |o_1 - o_2|$  is the absolute difference between the overall scores  $o$  and  
 1308  $o'$ . Note that also output distances are values between 0 and 1.

1309 Our scoring system is a function  $P : \mathcal{M}^5 \rightarrow \mathcal{M}$ . We will assume here that  $P$  is defined as  
 1310 the sum of five subscore systems, one for each of the five input marks, computing a value  
 1311 between 0 and 0.2. Then,

$$1312 \quad P((ed, ex, pe, in, sk)) := P_{\text{ed}}(ed) + P_{\text{ex}}(ex) + P_{\text{pe}}(pe) + P_{\text{in}}(in) + P_{\text{sk}}(sk).$$

1313 Let  $P_{\text{ed}}$ ,  $P_{\text{ex}}$ ,  $P_{\text{pe}}$  and  $P_{\text{in}}$  be defined according to the plot shown in Fig. 8a. With an  
 1314 increasing mark, these subscores increase up to an input mark of 0.8, whereafter the applicant  
 1315 becomes overqualified and the subscore slowly decreases.  $P_{\text{sk}}$  is depicted in Fig. 8b: The skill  
 1316 mark is less important, however a minimum amount of skills is required for the job. Hence,  
 1317 there is a jump of the skill score at an skill mark of roughly 0.19. Let John be an applicant  
 1318 with  $ed = ex = pe = in = 0.5$  and a skill mark of  $sk = 0.2$ , which maps to a skill score on  
 1319 the plateau after the jump. The subscores for education, experience, personality and mental  
 1320 ability mark are 0.12 each. The skill score computed for John is 0.05. Hence, John's overall  
 1321 score is  $P(\text{John}) = 4 \cdot 0.12 + 0.05 = 0.53$ . Let Synthia be a synthetic applicant with the  
 1322 same marks as John, except for the skill mark, which is 0.19 in Synthia's case. As depicted in  
 1323 Fig. 8b, the skill subscore for skill mark 0.19 is 0.02—Synthia is at the plateau right before  
 1324 the jump of the skill score. Her overall score is  $P(\text{Synthia}) = 4 \cdot 0.12 + 0.02 = 0.50$ . The  
 1325 input distance between John and Synthia is  $d_{\text{in}}(\text{John}, \text{Synthia}) = \sqrt{\frac{0.01^2}{5}} \approx 0.0045$  and the  
 1326 output distance is  $d_{\text{out}}(\text{John}, \text{Synthia}) = |0.53 - 0.5| = 0.03$ . It is easy to see that if we use  
 1327 Lipschitz-fairness, the Lipschitz constant  $L$  must be at least  $L = 6.7$  to allow the small jump  
 1328 in the skill subscore function. We argue that small jumps like those in the skill subscore  
 1329 are normal behaviour and, hence, fair. Assume for the remainder of this example that we use  
 1330 Lipschitz-fairness with  $L = 6.7$ .

1331 Consider now a slightly modified variant  $P'$  of  $P$ .  $P'$  is as  $P$  but uses a different subscore  
 1332 function  $P'_{\text{sk}}$  for the skill score. Fig. 8c shows the skill subscore function for  $P'$ .  $P'_{\text{sk}}$   
 1333 has a jump at skill mark 0.13 that is significantly larger than that in  $P_{\text{sk}}$ . We assume in  
 1334 this example that such a big jump is unfair. This assumption is warranted since, for many  
 1335 applications, such a small change in technical skills which has an immense impact on the  
 1336 skill subscore is not reasonable. Considering applicant John, his skill mark still maps to a

very high skill score of 0.19. Let Synclair be a third (potentially synthetic) applicant with  $\text{ed} = \text{ex} = \text{pe} = \text{in} = 0.5$  (as for John and Synthia) and  $\text{sk} = 0.13$ . Her skill mark maps to a very small skill score of 0.01. The overall scores are  $P'(\text{John}) = 4 \cdot 0.12 + 0.19 = 0.67$  and  $P'(\text{Synclair}) = 4 \cdot 0.12 + 0.01 = 0.49$ . The input distance is  $d_{\text{In}}(\text{John}, \text{Synclair}) = 0.0313$  and the output distance is  $d_{\text{Out}}(\text{John}, \text{Synclair}) = 0.18$ . Applying the Lipschitz condition to  $P'$  and  $d_{\text{In}}(\text{John}, \text{Synclair})$ , it is easy to see that  $d_{\text{Out}}(\text{John}, \text{Synclair})$  may become as large as 0.21. Hence,  $P'$  is classified as fair w.r.t. the Lipschitz condition. We see that a problem of the Lipschitz condition is that it is not possible to allow small jumps and at the same time disallow large jumps with equal increasing rate. This is because the distance of the inputs can only be used to multiply it with the Lipschitz constant.

$$f(d) = \begin{cases} 0.001 + 8d, & \text{for } d \in [0.0, 0.01] \\ 0.001 + 4d, & \text{for } d \in (0.01, 0.1] \\ 0.001 + 2d, & \text{for } d \in (0.1, 1.0] \end{cases}$$

Func-fairness is different in this regard. Function  $f$  receives the input distance and can freely define a bound on output distances based on the input distance. Indeed, the concrete  $f$  on the right overcomes the problem observed in the example. It uses the input distance for a case distinction on the magnitude of the input distance. For input distances up to 0.01,  $f$  effectively applies Lipschitz-fairness with  $L = 8$  to allow small jumps. For input distances between 0.01 and 0.1,  $f$  behaves like Lipschitz-fairness for  $L = 4$ , and for larger input distances, it enforces  $L = 2$ . In all cases we add 0.001 to the result to avoid  $f$  becoming zero (see footnote 7 on page 27 in the main paper). Applying func-fairness with  $\mathcal{C} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$  to  $P$ , the combination of John and Synthia (and hence the small jump of the skill score function) is not highlighted by FairnessAwareSystem, i.e., it is correctly detected as func-fair. Applied to  $P'$ , however, John and Synclair fall into the second case in the definition of  $f$ , but, as the emulated Lipschitz condition with  $L = 4$  is violated, FairnessAwareSystem likely finds a negative fairness score, i.e.,  $P'$  is not func-fair w.r.t. John. We remark that we propose this  $f$  for purely illustrative purposes. For real-world examples,  $f$  should be more sophisticated. Finding a suitable  $f$  can be a non-trivial task which hinges on various aspects that are crucial for the fairness evaluation in a given context. Clearly, the  $P$  and  $f$  provided in this illustration are toy examples that are probably inappropriate for real-world usage.

## A.1 Proofs

In this section, we will provide proofs for most of the propositions and theorems in the main paper. First, we show the correctness of the HyperSTL characterisations of robust cleanness and func-cleanness.

We first provide a lemma, which destructs the globally ( $\Box$ ) and weak until ( $\mathcal{W}$ ) operators such that the timing constraints encoded by these operators becomes explicit.

**Lemma 8** *Let  $\sigma : T \rightarrow X$  be a trace with  $T = \mathbb{N}$  or  $T = \mathbb{R}_{\geq 0}$  and let  $\phi$  and  $\psi$  be STL formulas. Then the following equivalences hold.*

1.  $\sigma, 0 \models \Box \phi$  if and only if  $\forall t \geq 0. \sigma, t \models \phi$ ,
2. if  $T = \mathbb{N}$ , then  $\sigma, 0 \models \phi \mathcal{W} \psi$  if and only if  $\forall t \geq 0. (\forall t' \leq t. \sigma, t' \models \neg \psi) \Rightarrow \sigma, t \models \phi$ .

**Proof** We prove the two statements separately.

1. Using the definition of the derived operators  $\Box$  and  $\Diamond$ , we get that  $\sigma, 0 \models \Box \phi$  holds if and only if  $\sigma, 0 \models \neg(\top \cup \neg \phi)$  holds. Using the (Boolean) semantics of STL, we get that

1378 this is equivalent to  $\neg(\exists t \geq 0. \sigma, t \models \neg\phi \wedge \forall t' < t. \sigma, t' \models \top)$ . After simple logical  
 1379 operations, we get that this is equivalent to  $\forall t \geq 0. \sigma, t \models \phi$  as required.

- 1380 2. Using , the definition of  $\mathcal{W}$ , the (Boolean) semantics of STL, and considering that  $\mathcal{T} = \mathbb{N}$ ,  
 1381 we get that  $\sigma, 0 \models \phi \mathcal{W} \psi$  if and only if  $\exists t \in \mathbb{N}. \sigma, t \models \psi \wedge \forall t' < t. \sigma, t' \models \phi$  or  
 1382  $\forall t \in \mathbb{N}. \sigma, t \models \phi$ . We denote this proposition as  $V$ . It is easy to see that the right operand  
 1383 of the equivalence to prove can be rewritten to  $\forall t \in \mathbb{N}. (\exists t' \leq t. \sigma, t' \models \psi) \vee \sigma, t \models \phi$ .  
 1384 We denote this proposition as  $W$  and must show that  $V \Rightarrow W$  and  $W \Rightarrow V$ . To prove  
 1385 that  $V$  implies  $W$ , we distinguish two cases.

- 1386 • For the first case, assume that the left operand of the disjunction in  $V$  holds, i.e.,  
 1387 there is some  $t \in \mathbb{N}$ , such that  $\sigma, t \models \psi \wedge \forall t' < t. \sigma, t' \models \phi$ . To show  $W$ , let  $t_0 \in \mathbb{N}$   
 1388 be arbitrary. If  $t \leq t_0$ , then there exists  $t' \leq t_0$  (namely  $t' = t$ ) such that  $\sigma, t' \models \psi$ ;  
 1389 hence  $W$  holds. If  $t > t_0$ , then we know from  $\forall t' < t. \sigma, t' \models \phi$  that  $\sigma, t_0 \models \phi$  is  
 1390 true; hence,  $W$  holds.  
 1391 • For the second case, assume that the right operand of the disjunction in  $V$  holds, i.e.,  
 1392  $\forall t \in \mathbb{N}. \sigma, t \models \phi$ . Then, obviously  $W$  holds.

1393 To prove that  $W$  implies  $V$ , let  $PV = \{t \in \mathbb{N} \mid \sigma, t \models \psi\}$  be the set of all time  
 1394 points at which  $\psi$  holds. If  $PV$  is the empty set, it follows immediately from  $W$  that  
 1395  $\forall t \in \mathbb{N}. \sigma, t \models \phi$  and that, hence,  $V$  holds. If  $PV$  is not empty, let  $t = \min PV$  be the  
 1396 smallest time in  $PV$  (the minimum always exists, because  $\mathcal{T} = \mathbb{N}$ ). Then, obviously,  
 1397  $\exists t \in \mathbb{N}. \sigma, t \models \psi$ . To show that  $V$  holds, it suffices to show that  $\forall t' < t. \sigma, t' \models \phi$ . This  
 1398 follows from  $W$ , because  $t$  is the smallest time at which  $\sigma, t \models \psi$  holds and, therefore,  
 1399 for every  $t' < t$  it does not hold that  $\sigma, t' \models \psi$ .

□

1401 Lemma 9 is specific to the HyperSTL formula (3); it converts it into a first-order logic  
 1402 formula.

1403 **Lemma 9** Let  $M \subseteq (\mathbb{N} \rightarrow X)$  be a discrete-time system and let  $Std \subseteq M$  be a set of standard  
 1404 traces. Also, let  $Std_\pi$  be a quantifier-free HyperSTL subformula, such that  $M, \{\pi := w\}, 0 \models Std_\pi$  if and only if  $w \in Std$ . Then,  $M, \emptyset, 0 \models \psi_{u\text{-rob}}$  if and only if

$$\begin{aligned} 1406 \forall w \in Std. \forall w' \in M. \exists w'' \in Std. (\forall t \geq 0. \text{eq}(w \downarrow_i[t], w'' \downarrow_i[t]) \leq 0) \wedge \\ 1407 \forall t \geq 0. (\forall t' \leq t. d_{\text{In}}(w'' \downarrow_i[t'], w' \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{\text{Out}}(w'' \downarrow_o[t], w' \downarrow_o[t]) - \kappa_o \leq 0. \end{aligned}$$

1408 **Proof** Using Lemma 8.1, Lemma 8.2, and Definition 6, we get that

$$\begin{aligned} 1409 M, \emptyset, 0 \models \forall \pi. \forall \pi'. \exists \pi''. Std_\pi \\ 1410 \rightarrow (Std_{\pi''} \wedge \square(\text{eq}(\pi \downarrow_i, \pi'' \downarrow_i) \leq 0) \wedge \\ 1411 ((d_{\text{Out}}(\pi'' \downarrow_o, \pi' \downarrow_o) - \kappa_o \leq 0) \mathcal{W} (d_{\text{In}}(\pi'' \downarrow_i, \pi' \downarrow_i) - \kappa_i > 0))) \end{aligned}$$

1412 holds if and only if

$$\begin{aligned} 1413 \forall w \in M. \forall w' \in M. \exists w'' \in M. (M, \Pi, 0 \models Std_\pi) \\ 1414 \rightarrow ((M, \Pi, 0 \models Std_{\pi''}) \wedge (\forall t \geq 0. (M, \Pi, t \models \text{eq}(\pi \downarrow_i, \pi'' \downarrow_i) \leq 0)) \wedge \\ 1415 (\forall t \geq 0. (\forall t' \leq t. (M, \Pi, t' \models \neg d_{\text{In}}(\pi'' \downarrow_i, \pi' \downarrow_i) - \kappa_i > 0)) \\ 1416 \Rightarrow (M, \Pi, t \models d_{\text{Out}}(\pi'' \downarrow_o, \pi' \downarrow_o) - \kappa_o \leq 0))) \end{aligned}$$

1417 holds for  $\Pi = \{\pi := w, \pi' := w', \pi'' := w''\}$ . Using the the constraint under which  $\text{Std}_\pi$  must  
 1418 be modelled, and by further applying Definition 6 and basic logical operations, we get that  
 1419 the above proposition is equivalent to

$$\begin{aligned} 1420 \quad & \forall w \in M. \forall w' \in M. \exists w'' \in M. w \in \text{Std} \\ 1421 \quad & \rightarrow \left( w'' \in \text{Std} \wedge (\forall t \geq 0. \text{eq}(w \downarrow_i[t], w'' \downarrow_i[t]) \leq 0) \wedge \right. \\ 1422 \quad & \left. (\forall t \geq 0. (\forall t' \leq t. d_{\text{In}}(w'' \downarrow_i[t'], w' \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{\text{Out}}(w'' \downarrow_o, w' \downarrow_o) - \kappa_o \leq 0) \right). \end{aligned}$$

1423 Finally, after carefully reordering premises, we get that the above holds if and only if

$$\begin{aligned} 1424 \quad & \forall w \in \text{Std}. \forall w' \in M. \exists w'' \in \text{Std}. (\forall t \geq 0. \text{eq}(w \downarrow_i[t], w'' \downarrow_i[t]) \leq 0) \wedge \\ 1425 \quad & \forall t \geq 0. (\forall t' \leq t. d_{\text{In}}(w'' \downarrow_i[t'], w' \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{\text{Out}}(w'' \downarrow_o, w' \downarrow_o) - \kappa_o \leq 0. \end{aligned}$$

1426  $\square$

1427 We omit the lemma analogue to Lemma 9 that reformulates formula (2) as a first-order  
 1428 characterisation. The proof for Proposition 3 further transforms the first-order characteri-  
 1429 sations of formulas (2) and (3) to show that they indeed match the definitions of l-robust  
 1430 cleanliness and u-robust cleanliness.

1431 **Proposition 3** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $\mathcal{C} = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, \kappa_i, \kappa_o \rangle$   
 1432 a contract or context for robust cleanliness with  $\text{Std} \subseteq L$ . Further, let  $\text{Std}_\pi$  be a quantifier-free  
 1433 HyperSTL subformula, such that  $L, \{\pi := w\}, 0 \models \text{Std}_\pi$  if and only if  $w \in \text{Std}$ . Then,  $L$  is  
 1434 l-robustly clean w.r.t.  $\mathcal{C}$  if and only if  $L, \emptyset, 0 \models \psi_{\text{l-rob}}$ , and  $L$  is u-robustly clean w.r.t.  $\mathcal{C}$  if  
 1435 and only if  $L, \emptyset, 0 \models \psi_{\text{u-rob}}$ .

1436 **Proof** We prove the correctness for l-robust cleanliness and u-robust cleanliness separately and  
 1437 begin with u-robust cleanliness. Using Lemma 9, we get that

$$\begin{aligned} 1438 \quad & L, \emptyset, 0 \models \forall \pi_1. \forall \pi_2. \exists \pi'_1. \text{Std}_{\pi_1} \\ 1439 \quad & \rightarrow \left( \text{Std}_{\pi'_1} \wedge \square(\text{eq}(\pi_1 \downarrow_i, \pi'_1 \downarrow_i) \leq 0) \wedge \right. \\ 1440 \quad & \left. (d_{\text{Out}}(\pi'_1 \downarrow_o, \pi_2 \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(\pi'_1 \downarrow_i, \pi_2 \downarrow_i) - \kappa_i > 0) \right) \end{aligned}$$

1441 holds if and only if

$$\begin{aligned} 1442 \quad & \forall w_1 \in \text{Std}. \forall w_2 \in L. \exists w'_1 \in \text{Std}. (\forall t \geq 0. \text{eq}(w_1 \downarrow_i[t], w'_1 \downarrow_i[t]) \leq 0) \wedge \\ 1443 \quad & \forall t \geq 0. (\forall t' \leq t. d_{\text{In}}(w'_1 \downarrow_i[t'], w_2 \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{\text{Out}}(w'_1 \downarrow_o, w_2 \downarrow_o) - \kappa_o \leq 0. \end{aligned}$$

1444 After applying simple logical operations and using that  $\text{eq}(i_1, i_2) = 0$  if and only if  $i_1 = i_2$ ,  
 1445 we get that this is equivalent to

$$\begin{aligned} 1446 \quad & \forall w_1 \in \text{Std}. \forall w_2 \in L. \exists w'_1 \in \text{Std} \text{ with } w_1 \downarrow_i = w'_1 \downarrow_i. \\ 1447 \quad & (\forall t \geq 0. (\forall t' \leq t. d_{\text{In}}(w'_1 \downarrow_i[t'], w_2 \downarrow_i[t']) \leq \kappa_i) \Rightarrow d_{\text{Out}}(w'_1 \downarrow_o[t], w_2 \downarrow_o[t]) \leq \kappa_o), \end{aligned}$$

1448 which, since we assumed  $\text{Std} \subseteq L$ , is equivalent to the definition of u-robust cleanliness for  
 1449 mixed-IO systems.

1450 The proof for l-robust cleanliness is analogue.  $\square$

1451 We recapitulate the proposition similar to Proposition 3 for func-cleanliness.

1452 **Proposition 4** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $\mathcal{C} = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$   
 1453 a contract or context for func-cleanliness with  $\text{Std} \subseteq L$ . Further, let  $\text{Std}_\pi$  be a quantifier-free  
 1454 HyperSTL subformula, such that  $L, \{\pi := w\}, 0 \models \text{Std}_\pi$  if and only if  $w \in \text{Std}$ . Then,  $L$  is  
 1455 l-func-clean w.r.t.  $\mathcal{C}$  if and only if  $L, \emptyset, 0 \models \psi_{\text{l-fun}}$ , and  $L$  is u-func-clean w.r.t.  $\mathcal{C}$  if and only  
 1456 if  $L, \emptyset, 0 \models \psi_{\text{u-fun}}$ .

1457 The proof for Proposition 4 is conceptually similar to the one for Proposition 3. The only  
 1458 difference is that instead of the reasoning about the  $\mathcal{W}$  construct, the globally enforced relation  
 1459 between output distances and the result of  $f$  must be proven equivalent in the HyperSTL  
 1460 formulas and func-cleanness. We omit the proofs here.

1461 **Correctness of STL characterisations**

1462 Next, we show the correctness of the STL characterisations, i.e., we will prove the correctness  
 1463 of Theorems 5 and 6. We do so by first establishing a connection between the HyperSTL  
 1464 and the STL characterisations.

1465 **Lemma 10** *Let  $M \subseteq (\mathbb{N} \rightarrow X)$  be a discrete-time system and let  $Std = \{w_1, \dots, w_c\} \subseteq M$   
 1466 be a finite set of standard traces. Also, let  $Std_\pi$  be a quantifier-free HyperSTL subformula,  
 1467 such that  $M, \{\pi := w\}, 0 \models Std_\pi$  if and only if  $w \in Std$ . Then,  $M, \emptyset, 0 \models \psi_{u\text{-rob}}$  if and only  
 1468 if  $(M \circ Std) \models \varphi_{u\text{-rob}}$  (with  $\varphi_{u\text{-rob}}$  from Theorem 5).*

1469 **Proof** Using Lemma 9 we get that

$$\begin{aligned} 1470 \quad M, \emptyset, 0 &\models \forall \pi'. \forall \pi''. \exists \pi'''. Std_\pi' \\ 1471 \quad &\rightarrow (Std_\pi''' \wedge \square(eq(\pi' \downarrow_i, \pi''' \downarrow_i) \leq 0) \wedge \\ 1472 \quad &\quad ((d_{out}(\pi''' \downarrow_o, \pi'' \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{in}(\pi''' \downarrow_i, \pi'' \downarrow_i) - \kappa_i > 0)) \end{aligned}$$

1473 holds if and only if

$$\begin{aligned} 1474 \quad \forall w' \in Std. \forall w'' \in M. \exists w''' \in Std. (\forall t \geq 0. eq(w' \downarrow_i[t], w''' \downarrow_i[t]) \leq 0) \wedge \\ 1475 \quad \forall t \geq 0. (\forall t' \leq t. d_{in}(w''' \downarrow_i[t'], w'' \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{out}(w''' \downarrow_o, w'' \downarrow_o) - \kappa_o \leq 0. \end{aligned}$$

1476 Since  $Std = \{w_1, \dots, w_c\}$ , we can replace the universal and existential quantifiers over  $Std$   
 1477 by a conjunction, respectively disjunction, over the standard traces [105]. We instantiate the  
 1478 universal quantifier for  $w''$  with  $w$  and get that

$$\begin{aligned} 1479 \quad \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} (\forall t \geq 0. eq(w_a \downarrow_i[t], w_b \downarrow_i[t]) \leq 0) \wedge \\ 1480 \quad \forall t \geq 0. (\forall t' \leq t. d_{in}(w_b \downarrow_i[t'], w \downarrow_i[t']) - \kappa_i \leq 0) \Rightarrow d_{out}(w_b \downarrow_o, w \downarrow_o) - \kappa_o \leq 0. \end{aligned}$$

1481 From the Boolean semantics of STL and by replacing all traces  $w$ , respectively  $w_k$ , by the  
 1482 corresponding  $w_+$ -projections, we get the equivalent proposition

$$\begin{aligned} 1483 \quad \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} (\forall t \geq 0. (w_+, t \models eq(w_a \downarrow_i, w_b \downarrow_i) \leq 0)) \wedge \\ 1484 \quad \forall t \geq 0. (\forall t' \leq t. (w_+, t' \models \neg d_{in}(w_b \downarrow_i, w \downarrow_i) - \kappa_i > 0)) \Rightarrow (w_+, t \models d_{out}(w_b \downarrow_o, w \downarrow_o) - \kappa_o \leq 0). \end{aligned}$$

1485 With the Boolean semantics of STL and Lemma 8.1 and 8.2 we get the equivalent statement  
 1486 that

$$\begin{aligned} 1487 \quad w_+, 0 \models \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} (\square(eq(w_a \downarrow_i, w_b \downarrow_i) \leq 0)) \wedge \\ 1488 \quad ((d_{out}(w_b \downarrow_o, w \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{in}(w_b \downarrow_i, w \downarrow_i) - \kappa_i > 0)). \end{aligned}$$

1489 □

1490 We are now able to prove Theorem 5.

1491 **Theorem 5** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $\mathcal{C} = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  a  
 1492 context for func-cleanness with finite standard behaviour  $\text{Std} = \{w_1, \dots, w_c\} \subseteq L$ . Then,  $L$   
 1493 is u-func-clean w.r.t.  $\mathcal{C}$  if and only if  $(L \circ \text{Std}) \models \varphi_{\text{u-fun}}$ , where

$$1494 \varphi_{\text{u-fun}} := \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} \left( \square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0) \wedge \right. \\ 1495 \quad \left. (\square(d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) - f(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i)) \leq 0)) \right).$$

1496 **Proof** The theorem follows from Proposition 3 and Lemma 10.  $\square$

1497 To prove Theorem 6, we establish the following lemma, which is analogue to Lemma 10,  
 1498 up to u-func-cleanness replacing u-robust cleanness.

1499 **Lemma 11** Let  $M \subseteq (\mathcal{T} \rightarrow X)$  be a system and let  $\text{Std} = \{w_1, \dots, w_c\} \subseteq M$  be a finite  
 1500 set of standard traces. Also, let  $\text{Std}_\pi$  be a quantifier-free HyperSTL subformula, such that  
 1501  $M, \{\pi:=w\}, 0 \models \text{Std}_\pi$  if and only if  $w \in \text{Std}$ . Then,  $M, \emptyset, 0 \models \psi_{\text{u-fun}}$  if and only if  
 1502  $(M \circ \text{Std}) \models \varphi_{\text{u-fun}}$  (with  $\varphi_{\text{u-fun}}$  from Theorem 6).

1503 The proof for Lemma 11 is, up to the different reasoning for  $\square(d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) -$   
 1504  $f(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i)) \leq 0)$  instead of  $(d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) - \kappa_o \leq 0) \mathcal{W}(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i) - \kappa_i > 0)$ , identical to that of Lemma 10. We omit it here.

1506 **Theorem 6** Let  $L \subseteq \mathbb{N} \rightarrow (\text{In} \cup \text{Out})$  be a mixed-IO system and  $\mathcal{C} = \langle \text{Std}, d_{\text{In}}, d_{\text{Out}}, f \rangle$  a  
 1507 context for func-cleanness with finite standard behaviour  $\text{Std} = \{w_1, \dots, w_c\} \subseteq L$ . Then,  $L$   
 1508 is u-func-clean w.r.t.  $\mathcal{C}$  if and only if  $(L \circ \text{Std}) \models \varphi_{\text{u-fun}}$ , where

$$1509 \varphi_{\text{u-fun}} := \bigwedge_{1 \leq a \leq c} \bigvee_{1 \leq b \leq c} \left( \square(\text{eq}(w_a \downarrow_i, w_b \downarrow_i) \leq 0) \wedge \right. \\ 1510 \quad \left. (\square(d_{\text{Out}}(w_b \downarrow_o, w \downarrow_o) - f(d_{\text{In}}(w_b \downarrow_i, w \downarrow_i)) \leq 0)) \right).$$

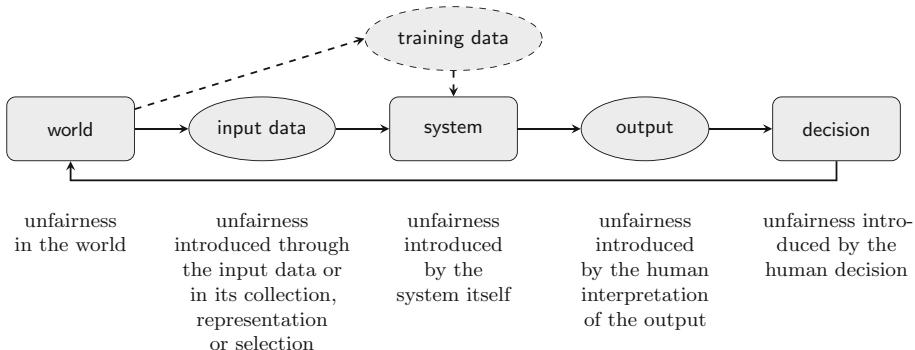
1511 **Proof** The theorem follows from Proposition 4 and Lemma 11.  $\square$

## 1512 Appendix B Fairness Pipeline

1513 As explained in Section 2 in the main paper, it is important to recognise that there are many  
 1514 sources of unfairness [8]. Section Appendix B shows a more detailed version of Fig. 5 in  
 1515 the main paper. Not every technical measure is able to detect every kind of unfairness and  
 1516 eliminating one source of unfairness might not be sufficient to eliminate all unfairness.

1517 **World** There can be unfairness in the world that leads to individuals  
 1518 already having worse (or better) starting conditions than others  
 1519 and subsequently have a lower (or higher) chance that the final  
 1520 decision is made in their favour. For example, an individual could  
 1521 be systematically excluded from certain societal resources (e.g.,  
 1522 girls who are excluded from education in Afghanistan under the  
 1523 Taliban) which puts these individuals at a disadvantage.

1524 **Input data** The input data or its collection, representation or selection could  
 1525 be problematic and lead to unfairness [139]. If, for example, crucial  
 1526 information is left out in the input data or data is aggregated  
 1527 in unsuitable ways, individuals could face an outcome that is  
 1528 unwarranted by the factual situation.



**Fig. 9** Sketch of different origins of unfairness in a decision process supported by a system; dashed elements are inapplicable to systems that are not learning-based

1529     **System (and training data)** The system itself can introduce new unfairness. Among other  
 1530     things, this can come about by erroneous algorithms or (in the  
 1531     case of a trained model) by problematic training data, e.g., if a  
 1532     certain group of individuals is not properly represented [120].

1533     **Output** The human decider can fail to interpret the output properly [138,  
 1534     140], which can lead to further unfairness. They could, for exam-  
 1535     ple, lack knowledge of the limitations of the system or fail to take  
 1536     into account that the system output is subject to some systematic  
 1537     uncertainty.

1538     **Decision** The human decider can make an unfair decision even in the face  
 1539     of a fair system output and an adequate interpretation thereof,  
 1540     for example if they have conscious or subconscious bias against  
 1541     certain groups [137].

1542     Unfairness in any part of the chain can arguably perpetuate or reinforce unfairness in the  
 1543     world.

1544     In the main paper, we propose a runtime monitoring technique that aims to uncover  
 1545     individual unfairness introduced by the system. By focusing on the system and its input-  
 1546     output relation only, we can say that the system is unfair without having to say anything  
 1547     about the degree of fairness with which an individual is treated in other respects in the  
 1548     decision process. It especially allows us to say that a system output is unfair, even though  
 1549     the outcome of the overall decision process is not. It may, for example, be that the system  
 1550     unfairness is ‘cancelled out’ by something else that is hidden from the system: an applicant  
 1551     with a stellar-looking CV might be treated unfairly by the system because of their age, but  
 1552     not hiring them is not unfair because they are known to have forged their diploma. Cases like  
 1553     this, however, do not make the unfairness introduced by the system any less problematic.

## 1554     Appendix C Legal Appendix

### 1555     C.1 EU anti-discrimination law

1556     Antidiscrimination is a principle deeply rooted in EU law. It is enshrined in Art. 21 of the  
 1557     Charter of Fundamental Rights (CFR) [47], which prohibits ‘[a]ny discrimination based on

any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation' as well as 'any discrimination based on grounds of nationality'. According to Art. 51 CFR, the addressees of this fundamental right are the EU and its institutions, bodies, offices and agencies as well as the Member States, insofar as they implement Union law. They are directly bound by Art. 21 above all in their legislative activities, but also in their executive and judicial measures. In contrast, private individuals are not directly bound by Art. 21 CFR, but they may be bound by regulations implementing this provision. However, according to recent European Court of Justice (ECJ) case law, Art. 21 CFR is directly applicable as a result of Directives, such as Directive 2000/78/EC [122] establishing a general framework for equal treatment in employment and occupation [45, §76]. Apart from this, while Art. 21 CFR stipulates a general prohibition of any unjustified discrimination, the more specific secondary legislation applicable to private actors only prohibits discrimination only in certain sensitive areas and only with regard to certain protected attributes. Correspondingly, private actors may not discriminate against certain persons—to name just a few—in employment relationships [122], in cases of abuse of a dominant market position [48, Art. 102] or also in so-called mass transactions, i.e., contracts that are typically concluded without regard to the person on comparable terms in a large number of cases [123]. In contrast, discriminating in other legal relationships or on other grounds such as local origin (as opposed to ethnic origin), or a person's financial situation is not generally prohibited. The rationale behind these 'discriminatory standards of anti-discrimination law' [57, 113, 126] is the principle of private (or personal) autonomy, and more specifically freedom of contract as one of its manifestations, which govern legal transactions between private individuals [75]. According to this principle, individuals are free to shape their legal relationships according to their own preferences and ideas, however irrational or socially unacceptable they may be. In essence, this also includes a right to discriminate against others. This freedom to autonomously form legal relations is only constrained where this is stipulated by anti-discrimination legislation for policy reasons.

When using an AI-system to recruit candidates, developers and deployers have to make sure that the system with its parameters comply with these legal requirements set by anti-discrimination law. This means in particular that the selection of the properties which a classifier should exhibit requires a positive normative choice and should not simply replicate biases present in the status quo [131]. However, the risks associated with deploying such systems in an HR context (such as a malfunctioning remaining undetected due to the system's opacity, a huge practical relevance of biased outputs due to the systems' scalability or the human operator's tendency of over-relying on the output produced by the AI system ('automation bias')), raise the question whether it can still be deemed normatively acceptable that the EU legal framework turns a blind eye on certain forms of discrimination. Furthermore, the principle of private autonomy as rationale for justifying the freedom to discriminate against others is only valid with regard to human's wilful actions, but not to algorithm-generated output. We are not advocating for abolishing the existing balance between private autonomy (freedom to contract) and prohibition to discriminate. So humans should still be permitted to differentiate on grounds that are not caught by anti-discrimination law. However, there is no reason to grant the 'right to discriminate' also to a non-human system that has merely "learned" this discrimination. In this respect, it seems justified to apply different standards for algorithms with regard to the prohibition of discrimination than for human decisions. With regard to an AI system's decision metrics, therefore, it should be considered to expand the secondary legal framework to include a broad prohibition of discrimination. This would not mean that all discrimination would be unlawful, since objectively justified unequal treatment

is, after all, permissible, but it would shift the focus to the question of objective justification [46]. Another legal challenge that will become even more pressing with the advent of technical decision systems is how to detect and prove prohibited discrimination. This is because the prohibition of discrimination resulting from various legal regulations in certain, especially sensitive, areas, such as human resources, presupposes that a difference in treatment is recognised in the first place. The recognition of discrimination is therefore not only in the interest of the decision-maker, who is threatened with sanctions in the event of a violation of the prohibition of discrimination. Rather, it is also essential for the discriminated party to prove the discrimination. For as far as a legal claim follows from a prohibited discrimination, the principle applies that the person who invokes the legal claim must prove the facts giving rise to the claim. Especially when complex algorithms are used, however, it is likely to be extremely difficult to prove corresponding circumstantial evidence. According to the case law of the ECJ, however, the burden of proof is reversed if the party who has *prima facie* been discriminated against would otherwise have no effective means of enforcing the prohibition of discrimination [42, 43]. Monitoring, as described here, would therefore be a suitable means of providing the ‘*prima facie*’ evidence necessary for shifting the burden of proof.

## 1624 C.2 Discrimination and the GDPR

1625 There has recently been discussion if and to which extent data protection law contains obligations for non-discriminating data processing or whether the scope of protection of data  
1626 protection law is thereby overstretched. There is no explicit prohibition of discrimination  
1627 in the General Data Protection Regulation (GDPR). According to Article 1 (2), however,  
1628 the GDPR is intended to protect the fundamental rights and freedoms of natural persons.  
1629 This is aimed in particular at their right to protection of personal data (Article 8 CFR), but  
1630 not exclusively so. Thus, the broad and non-restrictive reference to fundamental rights also  
1631 encompasses all other fundamental rights, including the right to non-discrimination (Article  
1632 21 CFR) [39]. This is reflected, for example, in the higher level of protection for data with an  
1633 increased potential for discrimination, the so-called special categories of personal data under  
1634 Article 9 GDPR. The GDPR can also be interpreted as granting a “preventive protection  
1635 against discrimination”, namely when discrimination is made impossible from the outset,  
1636 in that the data-processing agencies cannot gain knowledge of characteristics susceptible to  
1637 discrimination in the first place, i.e., when any respective data processing is forbidden [26].  
1638 Any processing of personal data must furthermore comply with the processing principles set  
1639 out in Article 5 GDPR, including the fairness principle (‘personal data shall be processed  
1640 fairly’) set out in Article 5(1)(a). While formerly transparency obligations were read into this  
1641 principle while the Data Protection Directive was into effect, the regulatory content of the  
1642 fairness principle is highly disputed since it was split off into a separate processing principle.  
1643 But due to the fact that discriminatory data processing can hardly be described as fair, a  
1644 prohibition of discrimination can be linked to the fairness principle [56, 77]. However, the  
1645 concrete scope of the fairness principle clearly goes beyond the understanding of fairness in  
1646 the context of technical systems on which this paper is based.

1647 Specifically for the HR context, there are discrimination-sensitive regulations in the  
1648 GDPR. Article 9 GDPR makes the processing of special categories of data, i.e., sensitive  
1649 data and data susceptible to discrimination, subject to particularly strict authorisation criteria,  
1650 which should in practice rarely be present in recruitment situations. On the one hand, process-  
1651 ing for recruitment purposes, i.e., prior to the establishment of an employment relationship,

is rarely necessary in order to exercise certain rights and obligations under employment law (Art. 9(2)(b) GDPR), and on the other hand, explicit consent (Art. 9(2)(a) GDPR) will often lack the necessary voluntariness due to the specifics of job application situations and the power imbalances inherent in them. The prohibition of processing sensitive data may be problematic in cases where the link to sensitive data is strictly necessary to detect discriminatory effects. For high-risk systems, Art. 10 V AI Regulation Proposal therefore provides for a new permissive clause: 'To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction, ... the providers of such systems may process special categories of personal data' while ensuring appropriate safeguards for the fundamental rights of natural persons.

With regard to the processing of non-sensitive personal data, however, the opening clause in Art 88(1) GDPR allows Member States to adopt more specific rules for processing for recruitment purposes, whereby, according to paragraph 2, suitable and specific measures must be ensured to safeguard the fundamental rights of the data subject. These requirements can be met by state-of-the-art monitoring tools. The national regulations cannot be discussed in depth here. For Germany, for example, Section 26 of the Federal Data Protection Act (BDSG) stipulates that personal data may only be processed for recruitment purposes if this is necessary, i.e., if the data processing is required for the decision on recruitment. In any case, data processing may not be necessary if the characteristics depicted in the data may not be taken into account in the hiring decision, for example due to anti-discrimination law [103].

## References

- Abbas H, Fainekos GE, Sankaranarayanan S et al (2013) Probabilistic temporal logic falsification of cyber-physical systems. ACM Trans Embed Comput Syst 12(2):95:1–95:30. <https://doi.org/10.1145/2465787.2465797>
- Alves WM, Rossi PH (1978) Who should get what? fairness judgments of the distribution of earnings. Am J Sociol 84(3):541–564
- Angwin J, Larson J, Mattu S, et al (2016) Machine bias. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Annapureddy YSR, Fainekos GE (2010) Ant colonies for temporal logic falsification of hybrid systems. In: IECON 2010—36th annual conference on IEEE industrial electronics society, pp 91–96, <https://doi.org/10.1109/IECON.2010.5675195>
- Arrieta AB, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
- Artistotle (1998a) The nicomachean ethics. Oxford worlds classics, Oxford University Press, Oxford, translation by W.D. Ross. Edition by John L. Ackrill, and James O. Urmson
- Artistotle (1998b) Politics. Oxford worlds classics, Oxford University Press, Oxford, translation by Ernest Barker. Edition by R. F. Stalley
- Baracas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671
- Barthe G, D'Argenio PR, Rezk T (2011) Secure information flow by self-composition. Math Struct Comput Sci 21(6):1207–1252. <https://doi.org/10.1017/S0960129511000193>
- Barthe G, D'Argenio PR, Finkbeiner B, et al (2016) Facets of software doping. In: Margaria T, Steffen B (eds) Leveraging applications of formal methods, verification and validation: discussion, dissemination, applications—7th international symposium, IsoLA 2016, Imperial, Corfu, Greece, October 10–14, 2016, Proceedings, Part II, pp 601–608, [https://doi.org/10.1007/978-3-319-47169-3\\_46](https://doi.org/10.1007/978-3-319-47169-3_46)
- Bathaei Y (2017) The artificial intelligence black box and the failure of intent and causation. Harvard J Law Tech 31:889
- Baum D, Baum K, Gros TP, et al (2023) XAI requirements in smart production processes: a case study. In: World conference on explainable artificial intelligence. Springer, pp 3–24
- Baum K (2016) What the hack is wrong with software doping? In: Margaria T, Steffen B (eds) Leveraging applications of formal methods, verification and validation: discussion, dissemination, applications—7th

- international symposium, ISoLA 2016, Imperial, Corfu, Greece, October 10-14, 2016, Proceedings, Part II, pp 633–647, [https://doi.org/10.1007/978-3-319-47169-3\\_49](https://doi.org/10.1007/978-3-319-47169-3_49),
14. Baum K, Mantel S, Schmidt E et al (2022) From responsibility to reason-giving explainable artificial intelligence. *Philos Tech* 35(1):12. <https://doi.org/10.1007/s13347-022-00510-w>
15. Biewer S (2023). Real driving emissions tests records. <https://doi.org/10.5281/zenodo.8058770>
16. Biewer S (2023b) Software doping—theory and detection. Dissertation (forthcoming)
17. Biewer S, Hermanns H (2022) On the detection of doped software by falsification. In: Johnsen EB, Wimmer M (eds) Fundamental approaches to software engineering—25th international conference, FASE 2022, Held as Part of the European joint conferences on theory and practice of software, ETAPS 2022, Munich, Germany, April 2–7, 2022, Proceedings, Lecture Notes in Computer Science, vol 13241. Springer, pp 71–91, [https://doi.org/10.1007/978-3-030-99429-7\\_4](https://doi.org/10.1007/978-3-030-99429-7_4),
18. Biewer S, D’Argenio PR, Hermanns H (2019) Doping tests for cyber-physical systems. In: Parker D, Wolf V (eds) Quantitative evaluation of systems, 16th international conference, QEST 2019, Glasgow, UK, September 10–12, 2019, proceedings, lecture notes in computer science, vol 11785. Springer, pp 313–331, [https://doi.org/10.1007/978-3-030-30281-8\\_18](https://doi.org/10.1007/978-3-030-30281-8_18),
19. Biewer S, D’Argenio PR, Hermanns H (2021) Doping tests for cyber-physical systems. *ACM Trans Model Comput Simul* 31(3):161–1627. <https://doi.org/10.1145/3449354>
20. Biewer S, Finkbeiner B, Hermanns H, et al (2021b) RTLOLA on board: testing real driving emissions on your phone. In: Groot JF, Larsen KG (eds) Tools and algorithms for the construction and analysis of systems—27th international conference, TACAS 2021, Held as Part of the European joint conferences on theory and practice of software, ETAPS 2021, Luxembourg City, Luxembourg, March 27 – April 1, 2021, Proceedings, Part II, Lecture Notes in Computer Science, vol 12652. Springer, pp 365–372, [https://doi.org/10.1007/978-3-030-72013-1\\_20](https://doi.org/10.1007/978-3-030-72013-1_20)
21. Biewer S, Dimitrova R, Fries M, et al (2022) Conformance relations and hyperproperties for doping detection in time and space. *Log Methods Comput Sci*. [https://doi.org/10.46298/lmcs-18\(1:14\)2022](https://doi.org/10.46298/lmcs-18(1:14)2022), [https://doi.org/10.46298/lmcs-18\(1:14\)2022](https://doi.org/10.46298/lmcs-18(1:14)2022)
22. Biewer S, Finkbeiner B, Hermanns H et al (2023) On the road with rtlola. *Int J Softw Tools Technol Transf* 25(2):205–218. <https://doi.org/10.1007/s10009-022-00689-5>
23. Binns R (2020) On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for computing machinery, New York, FAT\* ’20, pp 514–524, <https://doi.org/10.1145/3351095.3372864>,
24. Bloem R, Chatterjee K, Greimel K et al (2014) Synthesizing robust systems. *Acta Inf* 51(3–4):193–220. <https://doi.org/10.1007/s00236-013-0191-5>
25. Borgesius FJZ (2020) Strengthening legal protection against discrimination by algorithms and artificial intelligence. *Int J Human Rights* 24(10):1572–1593. <https://doi.org/10.1080/13642987.2020.1743976>
26. Buchner B (2020) DS-GVO Art. 1 Gegenstand und Ziele Rn. 14. In: Buchner JK (ed) Datenschutz-Grundverordnung, Bundesdatenschutzgesetz. C.H. Beck, Munich
27. Burke L (2020) The death and life of an admissions algorithm. <https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd>
28. Chazette L, Brunotte W, Speith T (2021) Exploring explainability: a definition, a model, and a knowledge catalogue. In: 2021 IEEE 29th international requirements engineering conference (RE), pp 197–208, <https://doi.org/10.1109/RE51729.2021.00025>
29. Chib S, Greenberg E (1995) Understanding the metropolis-hastings algorithm. *Am Stat* 49(4):327–335. <https://doi.org/10.1080/00031305.1995.10476177>
30. Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163. <https://doi.org/10.1089/big.2016.0047>
31. Clarkson MR, Finkbeiner B, Koleini M, et al (2014) Temporal logics for hyperproperties. In: Principles of security and trust—third international conference, POST 2014, Held as Part of the European joint conferences on theory and practice of software, ETAPS 2014, Grenoble, France, April 5–13, 2014, Proceedings, LNCS, vol 8414. Springer, pp 265–284, [https://doi.org/10.1007/978-3-642-54792-8\\_15](https://doi.org/10.1007/978-3-642-54792-8_15)
32. D’Argenio PR, Barthe G, Biewer S, et al (2017) Is your software on dope? - formal analysis of surreptitiously “enhanced” programs. In: Yang H (ed) Programming Languages and Systems - 26th European Symposium on Programming, ESOP 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22–29, 2017, Proceedings, Lecture Notes in Computer Science, vol 10201. Springer, pp 83–110, [https://doi.org/10.1007/978-3-662-54434-1\\_4](https://doi.org/10.1007/978-3-662-54434-1_4)
33. Donzé A, Ferrère T, Maler O (2013) Efficient robust monitoring for STL. In: Sharygina N, Veith H (eds) Computer aided verification—proceedings of 25th international conference, CAV 2013, Saint Petersburg, Russia, July 13–19, 2013. Lecture Notes in Computer Science, vol 8044. Springer, pp 264–279, [https://doi.org/10.1007/978-3-642-39799-8\\_19](https://doi.org/10.1007/978-3-642-39799-8_19)

- 1763 34. Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Sci Adv*  
1764 4(1):eao5580
- 1765 35. Dwork C, Hardt M, Pitassi T, et al (2012) Fairness through awareness. In: Proceedings of the 3rd  
1766 innovations in theoretical computer science conference, pp 214–226
- 1767 36. Dworkin R (1981) What is equality? Part 2: equality of resources. *Philos Public Aff* 10(4):283–345
- 1768 37. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Hum Factors* 37(1):32–  
1769 64. <https://doi.org/10.1518/001872095779049543>
- 1770 38. Endsley MR (2017) From here to autonomy: lessons learned from human-automation research. *Hum*  
1771 *Factors* 59(1):5–27. <https://doi.org/10.1177/0018720816681350>
- 1772 39. European Commission (2011) Proposal for a regulation of the European parliament and of the council  
1773 on the protection of individuals with regard to the processing of personal data and on the free movement  
1774 of such data (general data protection regulation) /\* com/2012/011 final. <https://eur-lex.europa.eu/legal->  
1775 [content/EN/TXT/?uri=celex%3A52012PC0011](#)
- 1776 40. European Commission (2021) Laying down harmonised rules on artificial intelligence (artificial intelli-  
1777 gence act) and amending certain union legislative acts (proposal for a regulation) no 0106/2021. <https://>  
1778 [eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206](#)
- 1779 41. European Commission (2023) Amendments adopted by the european parliament on 14 june 2023 on the  
1780 proposal for a regulation of the european parliament and of the council on laying down harmonised rules  
1781 on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. <https://>  
1782 [www.europarl.europa.eu/doceo/document/TA-9-2023-0236\\_EN.html](#)
- 1783 42. European Court of Justice (1993) C-127/92 - enderby ecli:eu:c:1993:859. <https://curia.europa.eu/juris/>  
1784 [liste.jsf?language=en&num=C-127/92](#)
- 1785 43. European Court of Justice (1995) C-400/93 - royal copenhagen ecli:eu:c:195:155. <https://curia.europa.>  
1786 [eu/juris/liste.jsf?language=en&num=C-400/93](#)
- 1787 44. European Court of Justice (2014) C-356/12 - glatzel ecli:eu:c:2014:350. <https://curia.europa.eu/juris/>  
1788 [liste.jsf?language=en&num=C-356/12](#)
- 1789 45. European Court of Justice (2018) C-414/16 - egenberger ecli:eu:c:2018:257. <https://curia.europa.eu/>  
1790 [juris/liste.jsf?language=en&num=C-414/16](#)
- 1791 46. European Parliament (2020) European parliament resolution of 20 october 2020 with recommendations  
1792 to the commission on a framework of ethical aspects of artificial intelligence, robotics and related  
1793 technologies. [https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275\\_EN.html](https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.html)
- 1794 47. European Union (2016a) Charter of fundamental rights of the european union. <https://eur-lex.europa.>  
1795 [eu/legal-content/EN/TXT/?uri=CELEX%3A12012P%2FTXT](#)
- 1796 48. European Union (2016b) Consolidated version of the treaty on the functioning of the european union.  
1797 <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12016ME%2FTXT>
- 1798 49. Fainekos GE, Pappas GJ (2009) Robustness of temporal logic specifications for continuous-time signals.  
1799 *Theor Comput Sci* 410(42):4262–4291. <https://doi.org/10.1016/j.tcs.2009.06.021>
- 1800 50. Ferrer X, Tv N, Such JM et al (2021) Bias and discrimination in AI: a cross-disciplinary perspective.  
1801 *IEEE Technol Soc Mag* 40(2):72–80. <https://doi.org/10.1109/MTS.2021.3056293>
- 1802 51. Finkbeiner B, Rabe MN, Sánchez C (2015) Algorithms for model checking HyperLTL and HyperCTL\*.  
1803 In: CAV 2015, LNCS, vol 9206. Springer, pp 30–48, [https://doi.org/10.1007/978-3-319-21690-4\\_3](https://doi.org/10.1007/978-3-319-21690-4_3)
- 1804 52. Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness: different  
1805 value systems require different mechanisms for fair decision making. *Commun ACM* 64(4):136–143.  
1806 <https://doi.org/10.1145/3433949>
- 1807 53. Gazda M, Mousavi MR (2020) Logical characterisation of hybrid conformance. In: Czumaj A, Dawar  
1808 A, Merelli E (eds) 47th international colloquium on automata, languages, and programming, ICALP  
1809 2020, July 8–11, 2020, Saarbrücken, Germany (Virtual Conference), LIPIcs, vol 168. Schloss Dagstuhl—  
1810 Leibniz-Zentrum für Informatik, pp 130:1–130:18, <https://doi.org/10.4230/LIPIcs.ICALP.2020.130>,
- 1811 54. Gunning D (2016) Explainable artificial intelligence (XAI) (darpa-baa-16-53). Tech. rep, Arlington, VA,  
1812 USA
- 1813 55. Guryan J, Charles KK (2013) taste-based or statistical discrimination: the economics of discrimination  
1814 returns to its roots. *Econ J* 123(572):F417–F432. <http://www.jstor.org/stable/42919257>
- 1815 56. Hacker P (2018) Teaching fairness to artificial intelligence: existing and novel strategies against algo-  
1816 rithmic discrimination under EU law. *Common Market Law Rev* (55):1143–1186. <https://ssrn.com/abstract=3164973>
- 1817 57. Hartmann F (2006) Diskriminierung durch Antidiskriminierungsrecht? Möglichkeiten und Grenzen  
1818 eines postkategorialen Diskriminierungsschutzes in der Europäischen Union. *EuZA - Europäische*  
1819 *Zeitschrift für Arbeitsrecht* p 24
- 1820

- 1821 58. Heaven WD (2020) Predictive policing algorithms are racist. They need to be dismantled. <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>
- 1822 59. High-Level Expert Group on Artificial Intelligence (2019) Ethics Guidelines for Trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- 1823 60. Hough LM, Oswald FL, Ployhart RE (2001) Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. *Int J Sel Assess* 9(1–2):152–194
- 1824 61. Ilvento C (2019) Metric learning for individual fairness. [arXiv:1906.00250](https://arxiv.org/abs/1906.00250)
- 1825 62. Jacovi A, Marasović A, Miller T, et al (2021) Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 624–635
- 1826 63. Jewson N, Mason D (1986) Modes of discrimination in the recruitment process: formalisation, fairness and efficiency. *Sociology* 20(1):43–63
- 1827 64. John PG, Vijayakeerthy D, Saha D (2020) Verifying individual fairness in machine learning models. In: Adams RP, Gogate V (eds) Proceedings of the thirty-sixth conference on uncertainty in artificial intelligence, UAI 2020, virtual online, August 3–6, 2020, Proceedings of machine learning research, vol 124. AUAI Press, pp 749–758. <http://proceedings.mlr.press/v124/george-john20a.html>
- 1828 65. Kästner L, Langer M, Lazar V, et al (2021) On the relation of trust and explainability: Why to engineer for trustworthiness. In: Yue T, Mirakhorli M (eds) 29th IEEE international requirements engineering conference workshops, RE 2021 workshops, Notre Dame, IN, USA, September 20–24, 2021. IEEE, pp 169–175, <https://doi.org/10.1109/REW53955.2021.00031>,
- 1829 66. Kim B, Khanna R, Koyejo O (2016) Examples are not enough, learn to criticize! criticism for interpretability. In: Proceedings of the 30th international conference on neural information processing systems. Curran Associates Inc., Red Hook, NIPS’16, pp 2288–2296
- 1830 67. Köhl MA, Hermanns H, Biewer S (2018) Efficient monitoring of real driving emissions. In: Colombo C, Leucker M (eds) Runtime Verification—Proceedings of 18th international conference, RV 2018, Limassol, Cyprus, November 10–13, 2018, Lecture Notes in Computer Science, vol 11237. Springer, pp 299–315, [https://doi.org/10.1007/978-3-030-03769-7\\_17](https://doi.org/10.1007/978-3-030-03769-7_17)
- 1831 68. Lai V, Tan C (2019) On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency, pp 29–38
- 1832 69. Langer M, Baum K, Hartmann K, et al (2021a) Explainability auditing for intelligent systems: a rationale for multi-disciplinary perspectives. In: Yue T, Mirakhorli M (eds) 29th IEEE international requirements engineering conference workshops, RE 2021 workshops, Notre Dame, IN, USA, September 20–24, 2021. IEEE, pp 164–168, <https://doi.org/10.1109/REW53955.2021.00030>,
- 1833 70. Langer M, Oster D, Speith T, et al (2021) What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell* 296(103):473. <https://doi.org/10.1016/j.artint.2021.103473>
- 1834 71. Langer M, Baum K, Schlicker N (2023) A signal detection perspective on error and unfairness detection as a critical aspect of human oversight of ai-based systems <https://doi.org/10.31234/osf.io/ke256>
- 1835 72. Larson J, Mattu S, Kirchner L, et al (2016) How we analyzed the COMPAS recidivism algorithm. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
- 1836 73. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. *Hum Factors* 46(1):50–80
- 1837 74. Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. *Entropy*. <https://doi.org/10.3390/e23010018>
- 1838 75. Looschelders D (2012) Diskriminierung und Schutz vor Diskriminierung im Privatrecht. *JZ - Juristen-Zeitung* p 105
- 1839 76. Maler O, Nickovic D (2004) Monitoring temporal properties of continuous signals. In: Lakhnech Y, Yovine S (eds) Formal techniques, modelling and analysis of timed and fault-tolerant systems, joint international conferences on formal modelling and analysis of timed systems, FORMATS 2004 and Formal Techniques in Real-Time and Fault-Tolerant Systems, FTRTFT 2004, Grenoble, France, September 22–24, 2004, Proceedings, Lecture Notes in Computer Science, vol 3253. Springer, pp 152–166, [https://doi.org/10.1007/978-3-540-30206-3\\_12](https://doi.org/10.1007/978-3-540-30206-3_12)
- 1840 77. Malgieri G (2020) What “fairness” means? A linguistic and contextual interpretation from the GDPR. In: FAT\* ’20: conference on fairness, accountability, and transparency, Barcelona, Spain, January 27–30, 2020. ACM, pp 154–166, <https://doi.org/10.1145/3351095.3372868>,
- 1841 78. Mathews M (2023) Are you ready for software-defined everything? *Wired*, <https://www.wired.com/insights/2013/05/are-you-ready-for-software-defined-everything/>, Accessed 23 June 2023
- 1842 79. Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 6(3):175–183. <https://doi.org/10.1007/s10676-004-3422-1>

80. Mecacci G, de Sio FS (2020) Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics Inf Technol* 22(2):103–115. <https://doi.org/10.1007/s10676-019-09519-w>
81. Mehrabi N, Morstatter F, Saxena N et al (2021) A survey on bias and fairness in machine learning. *ACM Comput Surv* 54(6):1–35
82. Meinke K, Sindhu MA (2011) Incremental learning-based testing for reactive systems. In: Gogolla M, Wolff B (eds) Tests and proofs—proceedings of 5th international conference, TAP@TOOLS 2011, Zurich, Switzerland, June 30–July 1, 2011. Lecture Notes in Computer Science, vol 6706. Springer, pp 134–151, [https://doi.org/10.1007/978-3-642-21768-5\\_11](https://doi.org/10.1007/978-3-642-21768-5_11)
83. Methnani L, Aler Tubella A, Dignum V et al (2021) Let me take over: variable autonomy for meaningful human control. *Front Artific Intell*. <https://doi.org/10.3389/frai.2021.737072>
84. Meurrens S (2021) The increasing role of AI in visa processing. <https://canadianimmigrant.ca/immigrate/immigration-law/the-increasing-role-of-ai-in-visa-processing>
85. Mittelstadt BD, Allo P, Taddeo M et al (2016) The ethics of algorithms: mapping the debate. *Big Data Soc* 3(2):2053951716679679. <https://doi.org/10.1177/2053951716679679>
86. Molnar C, Casalicchio G, Bischl B (2020) Interpretable machine learning—a brief history, state-of-the-art and challenges. In: Koprinska I, Kamp M, Appice A, et al (eds) ECML PKDD 2020 workshops—workshops of the European conference on machine learning and knowledge discovery in databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Communications in Computer and Information Science, vol 1323. Springer, pp 417–431, [https://doi.org/10.1007/978-3-030-65965-3\\_28](https://doi.org/10.1007/978-3-030-65965-3_28)
87. Mukherjee D, Yurochkin M, Banerjee M, et al (2020) Two simple ways to learn individual fairness metrics from data. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, proceedings of machine learning research, vol 119. PMLR, pp 7097–7107, <https://proceedings.mlr.press/v119/mukherjee20a.html>
88. Nghiem T, Sankaranarayanan S, Fainekos GE, et al (2010) Monte-carlo techniques for falsification of temporal properties of non-linear hybrid systems. In: Johansson KH, Yi W (eds) Proceedings of the 13th ACM international conference on hybrid systems: computation and control, HSCC 2010, Stockholm, Sweden, April 12–15, 2010. ACM, pp 211–220, <https://doi.org/10.1145/1755952.1755983>
89. Nguyen LV, Kapinski J, Jin X, et al (2017) Hyperproperties of real-valued signals. In: Talpin J, Derler P, Schneider K (eds) Proceedings of the 15th ACM–IEEE international conference on formal methods and models for system design, MEMOCODE 2017, Vienna, Austria, September 29 - October 02, 2017. ACM, pp 104–113, <https://doi.org/10.1145/3127041.3127058>
90. Noorman M (2020) Computing and Moral Responsibility. In: Zalta EN (ed) The stanford encyclopedia of philosophy, Spring, 2020th edn. Stanford University, Metaphysics Research Lab
91. Nunes I, Jannach D (2017) A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap Inter* 27(3):393–444
92. O’Neil C (2016a) How algorithms rule our working lives. <https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives>, Accessed 23 June 2023
93. O’Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Publishing Group, USA
94. Oracle (2019) AI in human resources: The time is now. <https://www.oracle.com/a/ocom/docs/applications/hcm/oracle-ai-in-hr-wp.pdf>
95. Organisation for Economic Co-operation and Development (OECD) (2021) Artificial intelligence, machine learning and big data in finance: opportunities, challenges and implications for policy makers. <https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf>
96. Pessach D, Shmueli E (2022) A review on fairness in machine learning. *ACM Comput Surv*. <https://doi.org/10.1145/3494672>
97. Pnueli A (1977) The temporal logic of programs. In: 18th annual symposium on foundations of computer science, Providence, Rhode Island, USA, 31 October–1 November 1977. IEEE Computer Society, pp 46–57, <https://doi.org/10.1109/SFCS.1977.32>
98. Rawls J (1985) Justice as fairness: Political not metaphysical. *Philos Public Affairs* 14(3):223–251. <http://www.jstor.org/stable/2265349>
99. Rawls J (1999) A theory of justice: Revised edition. Harvard university press
100. Rawls J (2001) Justice as fairness: a restatement. Harvard University Press
101. Ribeiro MT, Singh S, Guestrin C (2016a) Model-agnostic interpretability of machine learning. [abs/1606.05386](https://arxiv.org/abs/1606.05386), [arxiv:1606.05386](https://arxiv.org/pdf/1606.05386.pdf)
102. Ribeiro MT, Singh S, Guestrin C (2016b) “Why should I trust you?”: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery

- 1941 and data mining. Association for computing machinery, New York, KDD '16, pp 1135–1144, <https://doi.org/10.1145/2939672.2939778>.
- 1942 103. Riesenhuber K (2021) BDSG §26 Datenverarbeitung für Zwecke des Beschäftigungsverhältnisses Rn. 79f. In: Wolff SBA (ed) BeckOK Datenschutzrecht. C.H. Beck, Munich
- 1943 104. Rockafellar RT, Wets RJB (2009) Variational analysis, vol 317. Springer Science & Business Media
- 1944 105. Rosen KH, Krithivasan K (2012) Discrete mathematics and its applications: with combinatorics and 1945 graph theory. Tata McGraw-Hill Education
- 1946 106. Rowe T (2022) Can a risk of harm itself be a harm? *Analysis* 81(4):694–701. <https://doi.org/10.1093/analyse/anab033>
- 1947 107. Rubinstein RY (1981) Simulation and the Monte Carlo method. Wiley series in probability and mathematical statistics, Wiley <https://www.worldcat.org/oclc/07275104>
- 1948 108. Sankaranarayanan S, Fainekos G (2012) Falsification of temporal properties of hybrid systems using the cross-entropy method. In: Dang T, Mitchell IM (eds) Hybrid systems: computation and control (part of CPS Week 2012), HSCC'12, Beijing, China, April 17–19, 2012. ACM, pp 125–134, <https://doi.org/10.1145/2185632.2185653>,
- 1949 109. Sanneman L, Shah JA (2020) A situation awareness-based framework for design and evaluation of explainable AI. International workshop on explainable. Springer, Transparent Autonomous Agents and Multi-Agent Systems, pp 94–110
- 1950 110. Schlicker N, Langer M (2021) Towards warranted trust: a model on the relation between actual and 1951 perceived system trustworthiness. *Mensch Comput* 2021:325–329
- 1952 111. Schlicker N, Langer M, Ötting SK et al (2021) What to expect from opening up black boxes? comparing 1953 perceptions of justice between human and automated agents. *Comput Hum Behav* 122(106):837. <https://doi.org/10.1016/j.chb.2021.106837>
- 1954 112. Schlicker N, Uhde A, Baum K, et al (2022) Calibrated trust as a result of accurate trustworthiness 1955 assessment—introducing the trustworthiness assessment model. <https://doi.org/10.31234/osf.io/qhwvx>
- 1956 113. Schwab D (2006) Schranken der Vertragsfreiheit durch die Antidiskriminierungsrichtlinien und ihre 1957 Umsetzung in Deutschland. DNotZ—Deutsche Notar-Zeitschrift
- 1958 114. Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a 1959 philosophical account. *Frontiers in Robotics and AI* 5. <https://doi.org/10.3389/frobt.2018.00015>
- 1960 115. Smith E, Vogel H (2021) How your shadow credit score could decide whether you get an 1961 apartment. <https://www.propublica.org/article/how-your-shadow-credit-score-could-decide-whether-you-get-an-apartment>, Accessed 23 June 2023
- 1962 116. Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 1963 ACM conference on fairness, accountability, and transparency. Association for computing machinery, 1964 New York, FAccT '22, pp 2239–2250, <https://doi.org/10.1145/3531146.3534639>,
- 1965 117. Sterz S, Baum K, Lauber-Rönsberg A, et al (2021) Towards perspicuity requirements. In: Yue T, 1966 Mirakhorli M (eds) 29th IEEE international requirements engineering conference workshops, RE 2021 1967 Workshops, Notre Dame, IN, USA, September 20–24, 2021. IEEE, pp 159–163, <https://doi.org/10.1109/REW53955.2021.940029>,
- 1968 118. Tabuada P, Balkan A, Caliskan SY, et al (2012) Input-output robustness for discrete systems. In: Proceedings 1969 of the 12th International Conference on Embedded Software, EMSOFT 2012, part of the eighth 1970 embedded systems week, ESWeek 2012, Tampere, Finland, October 7–12, 2012. ACM, pp 217–226, 1971 <https://doi.org/10.1145/2380356.2380396>
- 1972 119. Talbert M (2019) Moral responsibility. In: Zalta EN (ed) The stanford encyclopedia of philosophy, 1973 Winter, 2019th edn. Stanford University, Metaphysics Research Lab
- 1974 120. Tay L, Woo SE, Hickman L et al (2022) A conceptual framework for investigating and mitigating 1975 machine-learning measurement bias (mlmb) in psychological assessment. *Adv Methods Pract Psychol Sci*. <https://doi.org/10.1177/25152459211061337>
- 1976 121. Technavio (2022) Software defined everything (SDE) market by end-user and geography—forecast 1977 and analysis 2022–2026. <https://www.technavio.com/report/software-defined-everything-sde-market-industry-analysis>, Accessed 23 June 2023
- 1978 122. The Council of the European Union (2000) Council directive 2000/78/EC of 27 november 2000 establishing 1979 a general framework for equal treatment in employment and occupation. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0078>
- 1980 123. The Council of the European Union (2004) Council directive 2004/113/EC of 13 december 2004 implementing 1981 the principle of equal treatment between men and women in the access to and supply of goods and services. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32004L0113>
- 1982 124. The European Parliament and the Council of the European Union (2017) Commission Regulation (EU) 1983 2017/1151. <http://data.europa.eu/eli/reg/2017/1151/oj>

125. Thüsing G (2013) European Labour Law, §3 Protection against discrimination. C.H. Beck
126. Thüsing G (2019) Das künftige Anti-Diskriminierungsrecht als Herausforderung für Wissenschaft und Praxis. ZfA - Zeitschrift für Arbeitsrecht p 241
127. Tutuiu M, Bonnel P, Ciuffo B et al (2015) Development of the world-wide harmonized light duty test cycle (WLTC) and a possible pathway for its introduction in the european legislation. Trans Res Part D Trans Environ 40(Suppl C):61–75. <https://doi.org/10.1016/j.trd.2015.07.011>
128. United Nations (2013) UN Vehicle Regulations—1958 Agreement, Revision 2, Addendum 100, Regulation No. 101, Revision 3—E/ECE/324/Rev.2/Add.100/Rev.3. [http://www.unece.org/trans/main/wp29regs101-120.html](http://www.unece.org/trans/main/wp29/wp29regs101-120.html)
129. United Nations Educational, Scientific and Cultural Organization (UNESCO) (2021) Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>
130. Volpatto M, Tretmans J (2015) Approximate active learning of nondeterministic input output transition systems. Electron Commun Eur Assoc Softw Sci Technol 72. <https://doi.org/10.14279/tuj.eceasst.72.1008>
131. Wachter S, Mittelstadt B, Russell C (2020) Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. W Va L Rev 123:735. <https://doi.org/10.2139/ssrn.3792772>
132. Washington State (2020) Certification of enrollment: engrossed substitute senate bill 6280 ('Washington State Facial Recognition Law'). <https://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Senate%20Passed%20Legislature/6280-S.PL.pdf?q=20210513071229>
133. Waters A, Miikkulainen R (2014) Grade: machine learning support for graduate admissions. AI Mag 35(1):64. <https://doi.org/10.1609/aimag.v35i1.2504>
134. Zehlike M, Yang K, Stoyanovich J (2021) Fairness in ranking: a survey. CoRR abs/2103.14000. <arxiv:2103.14000>,
135. Zemel R, Wu Y, Swersky K, et al (2013) Learning fair representations. In: International conference on machine learning, PMLR, pp 325–333
136. Ziegert JC, Hanges PJ (2005) Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. J Appl Psychol 90(3):553
137. Bertrand M, Mullainathan S (2004) Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. Am Econ Rev 94(4):991–1013
138. Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434
139. Lahoti P, Gummadi KP, Weikum G (2019) ifair: Learning individually fair data representations for algorithmic decision making. In: 2019 IEEE 35th international conference on data engineering (icde), IEEE, pp 1334–1345
140. Langer M, König CJ, Back C, et al (2022) Trust in artificial intelligence: comparing trust processes between human and automated trustees in light of unfair bias. J Bus Psychol

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Sebastian Biewer<sup>1</sup> · Kevin Baum<sup>1,2,3</sup> · Sarah Sterz<sup>1</sup> · Holger Hermanns<sup>1</sup> ·  
Sven Hetmank<sup>4</sup> · Markus Langer<sup>5</sup> · Anne Lauber-Rönsberg<sup>4</sup> · Franz Lehr<sup>4</sup>

Holger Hermanns  
hermanns@cs.uni-saarland.de

Sven Hetmank  
sven.hetmank@tu-dresden.de

Markus Langer  
markus.langer@uni-goettingen.de

2048 Anne Lauber-Rönsberg  
2049 anne.lauber@tu-dresden.de

- 2050 1 Department of Computer Science, Saarland University, Saarland Informatics Campus, Saarbrücken  
2051 66123, Germany
- 2052 2 Institute of Philosophy, Saarland University, Campus, Saarbrücken 66123, Germany
- 2053 3 Neuro-Mechanistic Modeling, German Research Center for Artificial Intelligence (DFKI),  
2054 Stuhlsatzenhausweg 3, Saarbrücken 66123, Germany
- 2055 4 IRGET, TU Dresden, Bergstraße 53, Dresden 01062, Germany
- 2056 5 Institute of Psychology, University of Göttingen, Goßlerstraße 14, Göttingen 37073, Germany

# Explainability as a Non-Functional Requirement

Maximilian A. Köhl\*, Dimitri Bohlender†, Kevin Baum\*, Markus Langer\*, Daniel Oster\* and Timo Speith\*

\*Saarland University, Saarbrücken, Germany

Email: mkoehl@cs.uni-saarland.de, {kevin.baum, markus.langer, daniel.oster, timo.speith}@uni-saarland.de

†RWTH Aachen University, Aachen, Germany

Email: bohlender@embedded.rwth-aachen.de

**Abstract**—Recent research efforts strive to aid in designing *explainable systems*. Nevertheless, a *systematic* and *overarching* approach to ensure explainability by design is still missing. Often it is not even clear what precisely is meant when demanding explainability. To address this challenge, we investigate the elicitation, specification, and verification of *explainability* as a *Non-Functional Requirement* (NFR) with the long-term vision of establishing a standardized certification process for the explainability of software-driven systems in tandem with appropriate development techniques.

In this work, we carve out different notions of explainability and high-level requirements people have in mind when demanding explainability, and sketch how explainability concerns may be approached in a hypothetical hiring scenario. We provide a conceptual analysis which unifies the different notions of explainability and the corresponding explainability demands.

**Index Terms**—explainable systems, requirements specification, requirements elicitation, terminology, certified explainability

## I. INTRODUCTION

The desire to sufficiently understand the systems we interact with is natural. If a person acts in an unexpected way, for instance comes late to an important appointment, we might ask for an explanation and be content with hearing about a traffic jam. In contrast, software-driven systems are becoming more and more opaque due to their ever increasing complexity and autonomy. Sometimes even domain experts and system engineers struggle to understand certain aspects of a system [1]. Systems with machine-learning based components in particular become hard to understand [2]. This development results in an increasing interest in *explainable systems*.

Explanations enable understanding and thereby foster trust and trustworthiness, justify actions and decisions, improve usability, help in locating sources of error, and can minimize the chance for human error. Particularly in “human in the loop” scenarios, in which humans have to make a decision based on a system’s recommendation, humans cannot reach an informed decision without having access to the system’s reasons for its recommendation. The assessment of a system’s allegedly erroneous behavior via an adequate explanation could resolve questions of responsibility and liability, e.g., whether the design was faulty and the manufacturer is to blame, or whether someone else is responsible.

A lack of explainability, on the other hand, not only gives rise to various moral, social, and legal problems [3], [4]. It further fuels distrust [5], diminishes user acceptance and satisfaction [6], and inhibits the adoption of new technologies. These problems have also been identified by legislators. The

European Union, for instance, debated about a general *Right to Explanation* [7] which is partly enshrined in certain regulations [8]. Furthermore, the EU High-Level Expert Group on AI proposed “Ethics Guidelines for Trustworthy AI,” in which they promote explainability as a crucial means for building trust in the decisions of software-driven systems [9].

Consequently, explainability needs to be taken into account during development in order to improve the quality of the target artifact and meet various regulatory requirements. Appropriate development techniques need to be established that guarantee a certain degree of explainability. Design decisions that impact the explainability of a system must not be taken by the developer implicitly, but explicitly specified as part of the design process. However, it is often unclear what precisely is meant when demanding explainability, and how it can be achieved by design. To the best of our knowledge there is no systematic and overarching approach to the explicit specification of explainability requirements, on how to take them into account during development, and how to evaluate whether an artifact indeed meets those requirements. This paper aims to be a starting point to address these issues.

### A. Contribution

We begin this paper with a brief discussion of research in the field of explainable artificial intelligence (XAI) and explainable systems in general. In Section II, we carve out different notions of explainability and high-level requirements people have in mind when demanding explainability. Further, we sketch how explainability concerns may be approached in a hypothetical hiring scenario in Section III. Based on the insights gained, we provide a conceptual analysis in Section IV, unifying the different notions of explainability and the corresponding explainability demands. The resulting notion provides a starting point for a systematic and overarching approach to explainability requirements. In Section V, we conclude by sketching our long-term vision of the establishment of a standardized certification process for explainability in tandem with appropriate development techniques.

## II. CHARTING THE FIELD

The demand for explainable systems is obvious in recent and ongoing research. A prominent example is the DARPA-funded *Explainable Artificial Intelligence* research project [1]. It acknowledges the lack of insight and knowledge gain in the context of current machine-learning based systems, and

aims to investigate (1) “how to produce more explainable models,” (2) “how to design the explanation interface,” and (3) “how to understand the psychological requirements for effective explanations.” Within the field of machine learning, the terminology surrounding explainability is neither uniform nor consistent [10]. Terms like “interpretability” [11], “scrutability” [12], and “explainable artificial intelligence” [13] target roughly the same endeavor, i.a. making the inner workings of systems more accessible and the outputs such as predictions or recommendations assessable. However, it is not clear what precisely is meant by “inner workings” or by “making accessible.” Even more pressing is that the same terms, e.g., “interpretability,” in different papers may refer to distinct notions. For instance, Lipton [2] mentions three different kinds of system transparency which alternately constitute interpretability.

Existing approaches aim at making given and usually non-interpretable systems explainable [14], [15]. System engineers attempt to generate explanations via feature importance [16] or explanation vectors [17]. Regarding document classification, Martens and Provost [18] propose linguistic explanations with bag-of-words features to make system recommendations more assessable for domain, e.g., legal or personnel, experts. To increase the transparency with respect to end users, comprehensible local approximations [19], counterfactuals [7], and contrastive explanations [20] come into play.

Research on explainable systems is carried out in pursuit of various goals, reflected in the different meanings attributed to “explainability.” There are roughly two categories of research: (A) research on how to adapt machine learning and other techniques to allow for a more thorough inspection and understanding for engineers who build such systems, and (B) research on how to enable users of such systems to understand them in relevant aspects. Here, deeper inspection and understanding of the system’s behavior for engineers is a prerequisite for making those systems understandable to users. Conflating those categories, explainability is concerned with enabling human understanding of various aspects of software-driven systems. In line with this, it is not obvious what policymakers or other stakeholders actually mean when they demand explainability and enshrine it in laws or guidelines. What shall be explainable to whom and how should it be evaluated whether an artifact indeed meets those requirements? These observations suggest that different target groups need different, context-sensitive explanations to be able to understand the relevant aspects of a particular system.

Studying different strains of research we find that, while different techniques for implementing explainability emerge, the concept itself, and in which context which techniques are appropriate, remains under-specified. All in all, there are many accounts with varying and partially overlapping goals. Despite the demand for explainability, there is no overarching consensus about what “explainability” means. Hence, a unified notion of explainability is needed. To take explainability into account during development, it needs to be specified more precisely, and knowledge about which techniques to apply in which case needs to be systematically collected.

### III. CASE STUDY: AUTOMATED HIRING SYSTEM

Let us now turn to a concrete example where explainability is required. Imagine the following scenario: a large organization tries to improve the efficiency of their hiring processes. In a meeting between the executive management, hiring managers, and employee representatives, they decide to task the IT department with developing and implementing a software system for trainee hiring [21]. In a first meeting with the IT department, the following functional requirements are identified: applicants shall apply through an online application system where they upload their CV. The system shall then automatically screen the applicants’ CVs, and provide the hiring managers with a ranking of applicants based on their estimated fit for a given position. The hiring managers can use this ranking as an additional source of information to screen the most promising candidates and, afterwards, decide which applicants proceed to the next stage of the hiring process. Based on this decision, applicants either receive a rejection letter or an invitation to the next stage of the selection process.

Among other requirements it is demanded that the system’s decisions, i.e., the ranking, “shall be explainable to the various stakeholders.” These are at least: applicants, hiring managers, the executive management, employee representatives, the legal department, and the engineers of the hiring system themselves [22]. All of these groups possess different background knowledge about the system and the hiring process, as well as different motivations within the hiring process. For instance, applicants only know that they upload their CV and want a fair hiring process [23]. The executive management wants a lean and effective process [24]. Employee representatives and the legal department may want to know on which features the system bases its ranking, as they want to have an unbiased selection process in order to prevent lawsuits [25]. To reach an informed decision based on the system’s ranking, the hiring managers demand reasons for why the system ranks applicants as it does. The mere ranking is not informative enough and insufficient for them to come to a justified decision.

In a meeting of the IT department a need for the following design choices becomes apparent: Roughly, the system could be implemented based on machine learning using existing data, by explicitly programming various criteria into the system, or by a combination of both. A purely machine-learning based approach using existing data will most likely introduce unfair biases and make the system hard to explain and reason about. On the contrary, a system using explicitly specified criteria will probably not perform as well and cost more, especially because the criteria need to be developed, and characteristics of a “good” applicant remain undefined without a solid job analysis. However, its reasoning would be explicit and easily explainable. With regard to the explainability demand of the hiring managers and the legal department, it is clear that some reasons need to be provided for why a certain ranking was produced.

At this stage, more precise requirements need to be elicited. What needs to be explainable to whom and what qualifies

as an explanation of what? Given precise and assessable explainability requirements, system engineers could explore the design space and determine appropriate development techniques in a more systematic and substantial way. For example, finding the right balance between machine-learning based components and explicit criteria such that the overall system becomes sufficiently explainable by design.

#### IV. EXPLAINABILITY REQUIREMENTS

We start with a conceptual analysis of *explainable systems* as an important first step towards a systematic and overarching approach for the elicitation and specification of explainability requirements. Intuitively, what makes certain aspects of a system explainable to the relevant stakeholders is access of the stakeholders to some kind of *explanation* for their aspects of interest. However, what is an explanation?

##### A. What is an Explanation?

Looking into literature one finds a broad variety of different approaches on how to spell out the concept of explanation. On the one side, there are rather technical notions of explanation [26], [14], [19] which are usually linked to causes. On the other side, there are more pragmatic notions which regard explanations as answers to certain questions, in particular “Why questions” [27], [28], [29]. Both approaches, however, do not exclude each other. An answer to a question can be an explanation precisely because it has the structure and qualities identified by technical accounts.

The need for an explanation originates in a lack of understanding of some phenomenon  $X$ , called *explanandum* in the philosophy of science [30]—however, not as a whole but with respect to some *aspect*  $Y$  of interest. Intuitively,  $Y$  encodes a question one may sensibly ask about  $X$ . For example, when applicant  $A$  is ranked higher than applicant  $B$ , one may ask the question: which qualities of  $A$  make  $A$  a better fit for the job than  $B$ ? Here, the explanandum  $X$  is the ranking produced by the system. When eliciting explainability requirements, it is crucial to precisely capture  $X$  and  $Y$  by interviewing the stakeholders. Questions like “Why does applicant  $A$  rank higher than applicant  $B$ ?” are ill-posed in that they are highly ambiguous. This question, for instance, can be answered simply by pointing out that  $A$  ranks higher than  $B$  because, according to the system,  $A$  is a better fit for the job. However, such an explanation will not be of much help for hiring managers.

For our purposes, we need a notion of explanation that targets a certain *kind* of stakeholder—an explanation for an engineer may not explain anything to a user. That is, we need a notion that enables generalization and abstracts from concrete individuals. Of course, referring to groups introduces imprecision as it is rarely possible to specify precise characteristics and skills of a certain group [31]. Nevertheless, it enables generalization and it is in fact a common technique to assume that users with specific skills interact with a system [32]. For our purposes, we aim to be able to express that something needs to be explainable to a particular group, viz. the *target group*  $G$  of an explanation. A characterization of the

concept of explanation which does not generalize and abstract from concrete individuals will not be very useful.

Still, a target group  $G$  may contain single agents who lack the required abilities or knowledge. To avoid such corner cases causing an explanation  $E$  to not qualify as such, even though it explains the aspect of interest to a significant part of  $G$ , we only require all *representatives*  $R$  of  $G$  to be content with the provided explanations. We presuppose that such representatives are equipped with the background knowledge and processing capabilities characteristic of the target group.

Furthermore, the *context* in which an explanation is provided matters. First, it does not only affect what needs to be explained. For example, before the hiring process, applicants might want to know which kind of information will be evaluated by the system and be interested in how to improve for their next selection process [33]. Second, different contexts may place constraints on the explanation generation process or the form of acceptable explanations. For example, to enable hiring managers to discuss the results in the context of a meeting, an explanation might have to fit on a single screen but still provide sufficient detail, or potentially be queryable at a reasonable latency to not hinder productivity.

Therefore, while some aspect  $Y$  may have an explanation that achieves the maximal depth of understanding, it might not be the best explanation in all contexts. In particular, if a context requires the explanation to be given in aural form, neither a detailed textual explanation nor a succinct visualization will suffice. Instead, for each context, an explanation must be found that maximizes the depth of understanding within the context’s constraints.

The notion we propose in the following is both *target-aware* and *context-aware*. What counts as an explanation of what for whom depends on the intended target group  $G$ , i.e., a certain kind of stakeholder, and the explanatory *context*  $C$ . Following insights of Achinstein [34] and Van Fraassen [35], we propose a pragmatic notion of “explanation for” in terms of understanding:

*Definition 1 (Explanation For):*  $E$  is an *explanation* of explanandum  $X$  with respect to aspect  $Y$  for target group  $G$ , in context  $C$ , if and only if the processing of  $E$  in context  $C$  by any representative<sup>1</sup>  $R$  of  $G$  makes  $R$  understand  $X$  with respect to  $Y$ .

Analyzing *explanation* in terms of *understanding* may not seem illuminating at first—however, as we argue, it is illuminating as it enables leveraging results from psychology and the cognitive sciences to assess whether something is really an explanation and how people react to different kinds of explanations [36], [37]. In particular, tying explainability to understanding eventually enables verification through studies conducted with the relevant stakeholders.

Note that our analysis is not supposed to conflict with, or replace, technical notions of *explanation*. In particular, an explanation may render  $X$  understandable with respect

<sup>1</sup>Note that we assume that  $R$  does not understand  $X$  with respect to  $Y$  yet. If they already understand  $X$  then nothing would *make* them understand.

to  $Y$  precisely because the explanation carries the relevant information and structure as identified by technical accounts.

Overall, the idea is to enable examination of explainability by measuring understanding, e.g., in psychological studies of whether the processing of certain explanations makes stakeholders understand the explanandum with respect to the relevant aspect in relevant contexts. Explainability is not a technical concept but tightly coupled to human understanding. As such, it is also a matter of degree and probability. In the following investigation, we mostly omit this quantitative nature of explanations. Future work, however, should investigate this more rigorously.

### B. Explainable Systems

What makes a system explainable with respect to a particular group in a certain context is the group member's access to explanations when in that context. To provide access to explanations, the latter need to be produced by something or someone. That which produces an explanation—what we here call the “means” of explanation—could be the system itself, another system, or even a human expert. The mere theoretical existence of some explanation is, however, not sufficient for a system to be explainable. We leverage the above characterization of explanation in order to specify what it takes for a system to be considered explainable:

*Definition 2 (Explainable System):* A system  $S$  is *explainable* by means  $M$  with respect to aspect  $Y$  of an explanandum<sup>2</sup>  $X$ , for target group  $G$  in context  $C$ , if and only if  $M$  is able to produce an  $E$  in context  $C$  such that  $E$  is an explanation of  $X$  with respect to  $Y$ , for  $G$  in  $C$ .

In general, a means  $M$  to produce an explanation of some aspect  $Y$  does not have to be part of the system  $S$  but may be provided by someone or something detached from  $S$ . Reconsidering the hiring example, explanations of the resulting ranking are dynamic and based on acquired data. It is natural to integrate the respective means directly into the system. However, in order to understand whether the system only considers applicant features that can legally be considered in hiring processes, applicants could also ask which criteria it considers. Such information about a system is static and already known at design time. As a result, the typical explanation of what a system's capabilities and features are is provided by human engineers in the form of documentation or a manual, which perfectly matches our characterization.

Just as the notion of explanation is not absolute but depends on an aspect of an explanandum and a target group in some context, a system is not just explainable *per se*, but only with respect to certain aspects, groups and contexts. Every unqualified use of the term “explainable” is under-specified.

### C. Explainability Requirements

With Definitions 1 and 2 in place, we can now capture explainability requirements. To meet the expectations of all stakeholders regarding explainability, we propose the following

<sup>2</sup>Here  $X$  is not an arbitrary explanandum but an explanandum related to  $S$ .

catalog of questions as a basis for elicitation of the requirements:

- 1) What are the relevant target groups  $G$ , e.g., engineers, end users, or lawyers, and which traits characterize each group's representatives  $R$ , e.g., specific background knowledge or cognitive capacities?
- 2) What are the explananda  $X$ , e.g., events or decisions?
- 3) Which aspects  $Y$  of the explananda  $X$  must be explained to which target group  $G$ , e.g., why is a decision justified, which causal chain of internal system events led up to it, why did some event  $e$  happen instead of event  $e'$ ?
- 4) In which context  $C$  may an aspect  $Y$  need explanation, and what are the implied constraints? For example, explanations might have to be aural in a driving situation.

Based on the answers to those questions, explainability requirements are then formulated using the following schema:

*Definition 3 (Explainability Requirement):* A system  $S$  must be explainable for target group  $G$  in context  $C$  with respect to aspect  $Y$  of explanandum  $X$ .

Conceptually, requiring a system to be explainable does not entail a specific function that the system must be capable of performing, but rather constrains how it may be implemented. Choosing certain development techniques may impede the ability to provide explanations with the desired qualities, or they may conflict with requirements like privacy, e.g., explanations may leak personal information about the applicants. In general, a trade-off between the degree of explainability and other goals must be made. In line with this and the tight coupling of explainability with human understanding we suggest to understand explainability requirements as *Non-Functional Requirements* (NFRs) of a specific kind that must be *satisfied* rather than satisfied [38].

In the following, we will illustrate how our understanding of explainability requirements facilitates the elicitation of requirements, and enables their consideration during development, in tandem with other NFRs.

*Softgoal Interdependency Graphs* (SIGs) [39] represent and record the software design and reasoning process as well as the relationship among multiple requirements in the context of NFRs. Here, the main requirements constitute the acyclic graph's top nodes which are iteratively refined into sub-softgoals, forming the graph's middle layer, and eventually flow into the bottom layer which links concrete development techniques, coined *operationalizations*, to the fine-grained softgoals.

When considering the explainability of a specific system, a central question is what this demand actually boils down to, i.e., which explanandum  $X$  must be explained. Given specified explainability requirements, these  $X$  are already identified, e.g., explainable decisions, and can be modeled as top-level softgoals of an SIG.

Decomposition and elicitation of sub-softgoals lie at the heart of building SIGs. Naturally, the requirement to make a system explainable can be decomposed guided by our concept of explainability. Given explainability requirements, the refinement of  $X$  with respect to the relevant aspects  $Y$

is already given, and can be improved by considering the groups  $G$  and contexts  $C$ . Incrementally refining the softgoals in this way facilitates systematic elicitation and decomposition of the overarching explainability softgoal since the scope of subgoals is increasingly constrained. In fact, related NFRs, like transparency of the code base [40], contribute to the overarching explainability softgoal and will occur as subgoals in the SIG.

However, explainability requirements may conflict with other softgoals such as performance, development cost, precision, or security. A less explainable system may be cheaper to build or could offer a higher performance. The SIG notion acknowledges this and offers *priorities* and *interdependency links* between softgoals as the central concepts to support decision making.

In an SIG, the explainability (sub-)softgoals will be placed among the other softgoals, such that conflicting ones can be linked and associated with a positive or negative contribution. Likewise, when possible operationalizations to realize explainability softgoals affect others in different ways, their contribution is tracked in the links. For example, when considering different machine-learning based operationalizations for classification systems, *neural networks* might increase performance and reduce development costs. However, they may lack interpretability, while the simpler *decision trees* may be found to have significantly better interpretability without sacrificing the other criteria. Embedding explainability requirements into SIGs makes such trade-offs explicit and enables recording of design decisions through further notation offered by SIGs.

To aid in the refinement, operationalization and conflict resolution processes, the *NFR Framework* [39] proposes to build knowledge bases, coined *catalogs*, that accumulate possible refinements and interdependencies considered in previous projects. Having appropriate catalogs at hand may help to alleviate that need, and simplify the construction of SIGs—in particular when developers need some source of domain knowledge before moving towards the actual operationalizations and the target artifact. To start the refinement catalog off, our explainability terminology induces several patterns, e.g., decomposition of explainability softgoals by target-groups.

Finally, based on Definition 1, an explainability requirement is met if and only if an explanation  $E$  is provided such that the processing of  $E$  in context  $C$  by any representative  $R$  of  $G$  makes  $R$  understand  $X$  with respect to  $Y$ . Mapping this pattern onto refinements in SIGs enables the decomposition of broad and abstract explainability softgoals, such as “the system must be explainable” or “decisions of the system must be explainable” down to fine-grained explainability requirements and softgoals. Explainability of the overall system is then satisfied by satisfying the resulting explainability sub-softgoals.

#### D. Assessing Understanding

In order to gain empirical confidence that a certain explanation is really understood by any representative, it might not be enough to provide a single representative with an explanation for a given explanandum and go through a checklist that assesses whether they understood the explanation. One

of the problems with such approaches is that an individual representative might still have idiosyncratic understanding of an issue. In addition, assessing understanding through self-report questionnaires tends to suffer from cognitive biases, e.g., when people overestimate their understanding. A more promising approach would be to choose a variety of representatives of a target group with different backgrounds, e.g., different age, gender, experience with a given problem, and provide them with explanations. The feedback from all these representatives could then be used to gain insights into the target group’s explanatory needs.

Furthermore, as the same explanation generally triggers different cognitive processes within people with different background and motivation [41], [42], it seems necessary to gain deeper insights into the representatives’ processing of explanations. For instance, one could use the think-aloud technique [43] trying to understand how people perceive a given explanation. After processing the explanation, the representatives could try to use self-explanation [44] to answer their own questions based on the explanation. This would show whether the explanation helped them to understand the issue. The representatives could then also be asked to try to transfer their new knowledge to a related issue [44]. This would help to evaluate whether the explanation not only helps them to understand a specific issue, but also enables people to transfer their new knowledge to new situations. These steps allow us to examine understanding within a person and are examples of how to assess whether a given system is explainable. The fields of cognitive science and education provide further ideas for insights into processes that generate understanding [45].

By relying on the concept of understanding, our overarching characterization makes explainability measurable, using established techniques from psychology and cognitive sciences. In any case, revealing that representatives were not able to follow an explanation and that it did not enhance their knowledge should lead to iterative processes to improve the overall explainability of the system.

## V. CONCLUSION

While explainability has become an important design desideratum it is under-specified what precisely is meant when demanding explainability. What shall be explainable to whom? How can an artifact be evaluated with respect to explainability requirements? How can explainability be achieved by design? In this paper, we briefly discussed various works in the area of explainable systems and presented a conceptual analysis which we used for the systematic specification and elicitation of explainability requirements.

Our long-term vision is to establish a standardized certification process in tandem with appropriate development techniques to achieve explainability by design. This paper is a starting point towards an overarching and systematic approach to explainability requirements. In future work, we intend to validate the proposed techniques in empirical studies, to develop explainability catalogs, and to identify potentially overlooked issues and improvements to our approach.

While we clarified what makes a system explainable and how explainability can be assessed empirically, further research is necessary on how to apply requirements and software engineering techniques to design explainable systems.

#### ACKNOWLEDGEMENTS

This work has received financial support by DFG grant 389792660 as part of TRR 248, by the ERC Advanced Investigators Grant 695614 (POWVER), and by the VolkswagenStiftung planning grant “Explainable Intelligent Systems.” It is based on discussions and work presented during the GI-Dagstuhl Seminar ES4CPS [46].

#### REFERENCES

- [1] “Broad agency announcement, explainable artificial intelligence (XAI), DARPA-BAA-16-53,” DARPA, Aug. 2016. [Online]. Available: <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>
- [2] Z. C. Lipton, “The mythos of model interpretability.” *Queue*, 2018.
- [3] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” *Big Data & Society*, 2016.
- [4] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Lütge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, “Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations,” *Minds and Machines*, 2018.
- [5] M. Lahijanian and M. Kwiatkowska, “Social trust: A major challenge for the future of autonomous systems,” in *2016 AAAI Fall Symposia, Arlington, Virginia, USA, November 17-19, 2016*, 2016.
- [6] L. R. Ye and P. E. Johnson, “The impact of explanation facilities on user acceptance of expert systems advice,” *MIS Quarterly*, 1995.
- [7] S. Wachter, B. D. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *arXiv preprint arXiv:1711.00399*, 2017.
- [8] The European Parliament and the Council of the European Union. (2016, April) Commission Regulation (EU) 2016/679. [Online]. Available: <http://data.europa.eu/eli/reg/2016/679/oj>
- [9] High-Level Expert Group on Artificial Intelligence. (2019) Ethics Guidelines for Trustworthy AI. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- [10] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, 2019.
- [11] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [12] J. Masthoff, N. Oren, K. van Deemter, and W. W. Vasconcelos, “Towards scrutible autonomous systems,” in *Symposium: Influencing People with Information*, 2012.
- [13] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.
- [14] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017.
- [15] O. Biran and C. Cotton, “Explanation and justification in machine learning: A survey,” in *IJCAI-17 workshop on explainable AI (XAI)*, 2017.
- [16] I. Kononenko, E. Štrumbelj, Z. Bosnić, D. Pevec, M. Kukar, and M. Robnik-Šikonja, “Explanation and reliability of individual predictions,” *Informatica*, vol. 37, no. 1, 2013.
- [17] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Mäller, “How to explain individual classification decisions,” *Journal of Machine Learning Research*, 2010.
- [18] D. Martens and F. Provost, “Explaining data-driven document classifications,” 2013.
- [19] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should I trust you?”. Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016.
- [20] T. Miller, “Contrastive explanation: A structural-model approach,” *arXiv preprint arXiv:1811.03163*, 2018.
- [21] A. Singh, C. Rose, K. Viswesvariah, V. Chenthamarakshan, and N. Kambohatla, “Prospect: A system for screening candidates for recruitment,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.
- [22] A. M. Ryan and N. T. Tippins, “Attracting and selecting: What psychological research tells us,” *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, 2004.
- [23] S. W. Gilliland, “The perceived fairness of selection systems: An organizational justice perspective,” *Academy of Management Review*, 1993.
- [24] D. E. Terpstra and E. J. Rozell, “The relationship of staffing practices to organizational level measures of performance,” *Personnel psychology*, 1993.
- [25] D. M. Truxillo, D. D. Steiner, and S. W. Gilliland, “The importance of organizational justice in personnel selection: Defining when selection fairness really matters,” *International Journal of Selection and Assessment*, 2004.
- [26] J. Y. Halpern, “Causes and explanations: A structural-model approach. part ii: Explanations,” *British Journal for the Philosophy of Science*, 2005.
- [27] B. C. Van Fraassen, *The Scientific Image*. Oxford University Press, 1980.
- [28] S. Bromberger, *On What We Know We Don’t Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. University of Chicago Press, 1992.
- [29] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, 2018.
- [30] J. Woodward, “Scientific explanation,” in *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2017.
- [31] A. G. Sutcliffe, S. Thew, and P. Jarvis, “Experience with user-centred requirements engineering,” *Requir. Eng.*, 2011.
- [32] A. Sutcliffe, *User-centred Requirements Engineering*. Springer Science & Business Media, 2012.
- [33] D. M. Truxillo, T. E. Bodner, M. Bertolino, T. N. Bauer, and C. A. Yonce, “Effects of explanations on applicant reactions: A meta-analytic review,” *International Journal of Selection and Assessment*, 2009.
- [34] P. Achinstein, *Evidence, Explanation, and Realism: Essays in Philosophy of Science*. Oxford University Press, 2010.
- [35] B. C. Van Fraassen, “The pragmatics of explanation,” *American Philosophical Quarterly*, 1977.
- [36] C. Bechlivianidis, D. A. Mullen, Lagnado, J. C. Zemla, and S. Sloman, “Concreteness and abstraction in everyday explanation,” *Psychonomical Bulletin & Review*, vol. 24, no. 5, pp. 1451 – 1464, 2017.
- [37] M. Langer, C. J. König, and A. Fitli, “Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection,” *Computers in Human Behavior*, vol. 81, pp. 19–30, 2018.
- [38] H. A. Simon, *The Sciences of the Artificial (3rd Ed.)*. MIT Press, 1996.
- [39] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, *Non-Functional Requirements in Software Engineering*. Springer, 2000.
- [40] L. M. Cysneiros, M. Raffi, and J. C. S. do Prado Leite, “Software transparency as a key requirement for self-driving cars,” in *26th IEEE International Requirements Engineering Conference, RE 2018, Banff, AB, Canada, August 20-24, 2018*, 2018.
- [41] T. Lombrozo, “The structure and function of explanations,” *Trends in cognitive sciences*, 2006.
- [42] L. J. Skitka, E. Mullen, T. Griffin, S. Hutchinson, and B. Chamberlin, “Dispositions, scripts, or motivated correction? understanding ideological differences in explanations for social problems.” *Journal of personality and social psychology*, 2002.
- [43] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. MIT Press, 1984.
- [44] M. T. Chi, N. De Leeuw, M.-H. Chiu, and C. LaVancher, “Eliciting self-explanations improves understanding,” *Cognitive science*, 1994.
- [45] R. White and R. Gunstone, *Probing Understanding*. Routledge, 2014.
- [46] J. Greenyer, M. Lochau, and T. Vogel, “Explainable software for cyber-physical systems (ES4CPS): report from the GI dagstuhl seminar 19023, january 06-11 2019, schloss dagstuhl,” *CoRR*, vol. abs/1904.11851, 2019.

## Academic References

---

### **Prof. Elijah Millgram**

Distinguished Professor in Philosophy

Philosophy Department  
215 Central Campus Dr, Room 467  
University of Utah  
Salt Lake City UT 84112  
USA

E-mail: [Lije.Millgram@m.cc.utah.edu](mailto:Lije.Millgram@m.cc.utah.edu), [elijah.millgram@gmail.com](mailto:elijah.millgram@gmail.com)

Web: <https://www.elijahmillgram.net/>

### **Prof. Dr. Marija Slavkovik**

Professor for Artificial Intelligence

Postboks 7802  
Universitetet i Bergen (UiB)  
5020 Bergen  
Norway

E-mail: [marija.slavkovik@uib.no](mailto:marija.slavkovik@uib.no)

Web: <https://slavkovik.com/>, <https://www.uib.no/en/persons/Marija.Slavkovik>

### **Prof. Dr. Susanne Mantel**

Professor for Practical Philosophy

Philosophisches Seminar  
Universität Heidelberg  
Schulgasse 6  
69117 Heidelberg

E-Mail: [susanne.mantel@uni-heidelberg.de](mailto:susanne.mantel@uni-heidelberg.de)

Web: [https://www.uni-heidelberg.de/fakultaeten/philosophie/philstem/personal/mantel\\_team.html](https://www.uni-heidelberg.de/fakultaeten/philosophie/philstem/personal/mantel_team.html)

### **Prof. Dr. Markus Langer**

Professor of Work and Organizational Psychology

Albert-Ludwigs-Universität Freiburg  
Institut für Psychologie  
Engelbergerstraße 41  
79085 Freiburg

E-Mail: [Markus.Langer@psychologie.uni-freiburg.de](mailto:Markus.Langer@psychologie.uni-freiburg.de)

Web: <https://www.psychologie.uni-freiburg.de/Members/langer/langer.html>

### **Prof. Dr. Anne Lauber-Rönsberg**

Professor of Civil Law, Intellectual Property Law, in particular Copyright Law, as well as Media and Data Protection Law

Institut für Internationales Recht, Geistiges Eigentum  
und Technikrecht (IRGET)  
Philosophische Fakultät der TU Dresden  
01062 Dresden

E-Mail: [office.lauber-roensberg@tu-dresden.de](mailto:office.lauber-roensberg@tu-dresden.de)

Web: <https://tu-dresden.de/gsw/phil/irget/jfbim13/die-professur/prof-dr-anne-lauber-roensberg>

## **Habilitation Equivalence**

---

As to qualifications that go beyond a Ph.D., I suggest considering the following features of my academic career as equivalent to a habilitation. The corresponding publications and activities are listed in the CV and the list of publications, where they are marked with a »⊗«.

### **Published Research on Topics Unrelated to That of the Dissertation**

---

The CV lists 18 research articles, all either published or accepted for publication. They all concern the ethics of digitalization in various ways and, thus, are unrelated to that of the doctoral dissertation. The corpus comprises 297 pages and can be accessed in full here: <http://kevinbaum.de/assets/zip/Corpus.zip>. I am listed as a principal or senior author for 10 of the papers. At the time of application, my *Hirsch Index* is 10 with over 725 citations, 333 last year alone.

### **Responsibilities Equal to and Beyond Those of a Junior Professor**

---

As **head of CERTAIN**, my responsibilities include:

- the *strategic and executive direction* of the center,
- the *initiation of research collaborations* with regional, national, and international partners from *industry and academia*,
- and the *acquisition of further satellite third-party projects*.

Further, I *independently* (»eigenverantwortlich«) determine research priorities and am responsible for the personnel of what is now a team of eight (which is constantly growing).

As **deputy head of the research department NMM**, I share responsibilities for

- the *strategic positioning* of the department,
- *personnel decisions*,
- and the *supervision of Master's and Bachelor's students as well as research assistants*, including the thematic design and (co-)supervision of *Research Immersion Labs*.

Further, I am *responsible for coordinating* with other research departments and *initiating collaborations as well as industry and third-party funded projects*. I also *carry out independent research* (in the sense of a traditional post-doc or junior research group position).

As **manager of the research department**, I am responsible for *administration and scientific management* as well as *accounting and general organizational issues*, from *resource allocation* and working time tracking to training opportunities and team event.

## **Teaching & Supervising**

---

My teaching experience now includes 24 courses, including:

- From 2015 to 2024, 12 (pro-)seminars on philosophical ethics, among others at the *Bucerius Law School* in Hamburg and the *Berlin University of the Arts*, all organized and held independently (»eigenverantwortlich«).
- From 2016 to 2023, 8 editions of the interdisciplinary lectures *Ethics for Nerds* and 2 derivatives on the ethics of digitization for the LLM *Informationstechnologie und Recht*. *Ethics for Nerds* received the »*Hochschulperle 2020*« of the *Stifterverband*, an award that honors particularly innovative and exemplary projects at universities throughout Germany.
- 2 further seminars on philosophical logic and applied AI for the societal good.

In addition, a signature course on practical AI ethics for the *KI Campus* is currently being finalized.

I have independently supervised and assessed 2 Bachelor's theses in Philosophy. I am currently supervising a Master's thesis in Computer Science. I have also been an advisor for 5 more Bachelor's, Master's, and state examination theses and 1 dissertation. I have also planned and carried out one *Research Immersion Lab* and am currently initiating another.