Analyzing Epistemic Injustice

Joris HULSTIJN [a,1], Huimin DONG [a] and Réka MARKOVICH [a]
[a] *University of Luxembourg, Esch-sur-Alzette, Luxembourg*

Abstract. Cases of bias and unfair decisions in automated decision-making are heavily discussed. When unfair decision outcomes can be attributed to an unjust difference in the knowledge of various groups of subjects, we can speak of *epistemic injustice* (Fricker). In this paper, we analyze various kinds of epistemic injustice, such as testimonial, hermeneutical, distributional, and content-based epistemic injustice, and show how they can be conceptualized. We then provide a formalization of the difference in group knowledge, in a version of epistemic logic. After that, we discuss a case of a badly designed information system for government decision-making: Toeslagenffaire (Netherlands). We analyze key observations from the case and show that they constitute a form of epistemic injustice.

Keywords. social epistemology, injustice, bias, epistemic logic

1. Introduction

Automated decision-making systems (will) make decisions that matter. An important concern about applications of AI is bias in decision-making [18, 25]. Systems, as well as humans who are supported by (or provide the data to such) systems, do make decisions that are unfair or unjustified for certain groups, relative to other groups. Some groups in society do not know the procedures or are unable to fill out the required forms [23]. Therefore, these groups are treated unfairly. For example, some groups are relatively more likely to get into trouble with the tax office [10, 19]. How can we analyze such cases? In general, errors in automated decision-making may occur because (i) the algorithm or decision rules are biased, (ii) the data set on which the system was trained is biased, or (iii) the human operator who should counterbalance possible system bias, is not supported to do this difficult task [18].

Social epistemology investigates the epistemic aspects of social interactions [17]. Scholars have proposed to analyze some of the above-mentioned cases in terms of *epistemic injustice* [15]. The term has two parts: (i) *injustice*: there is a moral wrong or a legal right that is violated (ii) *epistemic*: the wrong is based on a difference in the knowledge or information that is available to some groups in society relative to other groups [15, 12, 13, 20]. In this paper, we provide an initial formal conceptual analysis of the epistemic part, showing that in some cases (types) the injustice is due to the decision-maker's epistemic state (regarding the subject group's knowledge).

The discussion and debate about epistemic injustice are part of a wider trend combining ethics and epistemology, for instance in business ethics [12] and medical ethics

[1] Corresponding Author: Universitty of Luxembourg, Esch-sur-Alzette, E-mail:joris.hulstijn@uni.lu

[6]. Here we will look at another application domain: government (public administration) decision-making, see e.g. [23]. As a case, we use observations from a scandal about errors in government decision-making: the 'Toeslagen affaire' in the Netherlands [10]. We briefly compare it with the RoboDebt case in Australia [19]. The aim of the paper is to:

(1) identify the main types of epistemic injustice from the literature, and provide a possible explanation for the mechanisms that give rise to epistemic injustice
(2) develop an epistemic logic to specify the difference in knowledge between groups.
(3) Explain observations from the case [10], to establish whether those constitute a form of epistemic injustice.

The paper is structured as follows. In Section 2 we analyze the notion of epistemic injustice. In Section 3 we formalize the definition in an epistemic logic. Section 5 describes the cases, and provides observations that illustrate epistemic injustice. The paper ends with conclusions and suggestions for further research.

2. What is Epistemic Injustice?

Roughly, epistemic injustice is a form of injustice related to knowledge. More precisely "Epistemic injustice refers to those forms of unfair treatment that relate to issues of knowledge, understanding, and participation in communicative practices." [22, preface].

The notions of *justice* and *injustice* have been widely discussed in moral and legal philosophy. We do not recall this literature here, but focus only on epistemic injustice, based on how it is handled in social epistemology. In the context of social epistemology [22], we look at differences between groups in society. So, injustices are studied that can be attributed to a (lack of) knowledge in one group, relative to another. As we will point out, in several cases, not the actual knowledge of one agent or group is what is relevant, but the beliefs—or prejudices—of others regarding it.

Types of epistemic injustice Fricker [15] discusses three types of epistemic injustice.

The first type is the obvious case where one group has less (access to) knowledge than another (privileged) group, leading to unfair treatment. For lack of a better word, we will call this *distributional injustice*, because it is based on an unequal distribution of knowledge or education [16, p 1317]. We should emphasize however, that not all cases of unequal distribution of knowledge constitute an injustice. A large part of society is based on merit and education. People who study and know more, get more opportunities. It only becomes an injustice, when access to knowledge is withheld from certain groups. This may be, for example, when documents are not made available in a minority language.

The second type, *testimonial injustice*, is a form of unfairness related to the trust in someone's capacity as a knower, for example, as an expert witness in a trial. An injustice of this kind can occur when someone is not believed or even ignored, because of group properties like their sex, sexuality, gender presentation, race, disability, or more generally, because of their identity [15]. [2]

[2]When Marilyn Von Savant, the person having the highest recorded IQ in the world, provided the *correct* answer to the Monty Hall problem in the column of Parade Magazine, tens of thousands of people (including many mathematicians and other academics) reacted publicly rejecting *harshly* what Von Savant said. Most of

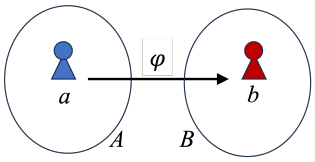

Figure 1. General setting of knowledge transfer φ between speaker a and addressee b, from groups A and B

The third type, *hermeneutical injustice* is a form of unfairness related to how people interpret and understand their situation, and their ability to formulate believable statements about that situation. The term 'hermeneutic' comes from the Greek word for interpreter. After all, hermeneutics is the philosophical discipline that is concerned with interpretation or understanding. An example given by Fricker, is that before the 1970s, victims of sexual harassment had trouble describing in court the behavior of which they were the victims, because the concept had not yet been articulated. In particular, legal procedures demanded physical evidence of abuse, which was hard to obtain.

A fourth version, *content-based epistemic injustice*, was suggested much later [14]. Here the unfair treatment is not based on characteristics of the group, but rather on the content of what they say. A particular type of message is ignored or mistaken, for example, because it doesn't align with the consensus opinion in the dominant group.

Setting Consider the following situation (Figure 1) There are two groups: A and B, whose members speak the same language, but have different knowledge, a different ontology to conceptualize the world, and a different terminology, to express themselves. Suppose speaker $a \in A$ must make a statement φ to addressee $b \in B$, because as part of his tasks b must make an important decision in the interest of a, for example, grant a subsidy. Formally, the decision depends on b accepting a's statement φ as true. Statement φ is typically supported by one or more documents of written evidence (letters; financial statements; tax returns etc). For example, if b doesn't believe that a is eligible for a subsidy, because a didn't provide proof of residence in the municipality in which the subsidy is claimed, b will not grant the subsidy.

Speech act theory We will now sketch a simplified model of the situation in Figure 1, based on speech act theory [2, 24]. In Table 1 we analyze the utterance in several layers: the form (syntax), the content (semantics), and the function (pragmatics). At each layer a performs various acts: (1a) locutionary: sending a statement message, (2a) illocutionary: conceptualizing and encoding the statement, and (3a) perlocutionary: making a statement

the reactions just considered her answer *unimaginable* to be correct, which falls into the category of content-based injustice (C4), but several reflected upon her being a woman as a reason for being wrong, which is a case of testimonial injustice (C2). See: https://priceonomics.com/the-time-everyone-corrected-the-worlds-smartest/

	a		b
1. locutionary:	sending a statement	–	receiving the statement
2. illocutionary:	conceptualizing and encoding the meaning	–	understanding and decoding the meaning
3. perlocutionary:	making a valid statement	–	accepting the statement as valid

Table 1. Model of information exchange as joint action at various linguistic levels [2, 9]

as part of a valid decision request. These acts from a are complemented by corresponding acts by b, at each level: (1b) receiving the statement, (2b) understanding the meaning of the statement, and (3b) accepting the statement as valid. This model reflects the idea of Clark [9] that communication is essentially a joint action by speaker and addressee, at various linguistic levels.

If we say "b doesn't believe a's statement φ", this may have several possible reasons. Given this model, there are in fact six potential points, in which something may have gone wrong. Now the four types of epistemic injustice can be positioned in this model.

C1 **Distributional**. Members of A lack knowledge φ, which members of B do have. Knowledge of φ is necessary to obtain some benefit ψ. So a cannot make a statement that φ (1a) that is received by b (1b), by lack of preparatory conditions.

C2 **Testimonial**. Members of A are in general not seen as trustworthy, about topics related to φ. Therefore, b doesn't accept statement φ as true (3b), by a supposed lack of sincerity of a.

C3 **Hermeneutical**. Members of A lack knowledge, to conceptualize and encode the intended meaning φ in a statement, that b will accept. (2a)

C4 **Content-based**. Members of B share a consensus that φ is false. Therefore, content φ will not be understood by b (2b), or a's statement φ will not be accepted by b to be true (3b).

For content-based epistemic injustice, there are two possible explanations. (i) confirmation bias. φ may not be accepted as true, in order to protect the group consensus (3b). If an individual b would accept φ, he would threaten the group's consensus or would place himself outside of the group. (ii) cognitive dissonance. Statement φ is so far removed from what is considered normal, that for b it takes much more effort to process and accept it, than to reject it. For example, b lacks the ontology to understand the problems of a (2b). This seems to be the counterpart of hermeneutical injustice for the addressee.

Sometimes, these types of epistemic injustice strengthen each other. For example, suppose b believes that φ must be false because it goes against the consensus (content-based). So for b, the reason that a makes a false statement φ can only be that a seeks to gain from it. So b doubts the sincerity of a, which is a case of testimonial injustice.

The world must correspond to system information.

Example: filing a complaint The two cases we will discuss in Section 5 below [10, 19], are both prescriptive systems: they have a word-to-world direction of fit. In case of a conflict about the contents of the system, the subjects of the decision are at a disadvantage. To file a complaint, they have to substantiate the claim by evidence, but, by the nature of the system, often there is no credible source to testify their version of reality. That means that a conflict becomes a game of trust. Now consider a subject to an automated decision, who feels she is being wronged. That means she must file a complaint. However, in both cases [10, 19], complaints were ignored or rejected without motivation.

Based on [15, 14] we identify four categories of epistemic injustice, that can be used to explain the social mechanisms why complaints were ignored. In some cases, more categories apply. For example, a conflict of opinion (C4) may further reduce the perceived sincerity of the group (C2).

C1 *Distributional injustice*: subjects who complain about a system error, lack technical knowledge about the system, lack legal knowledge of the grounds for the

decision, lack knowledge of the procedures to file a complaint, or do not know people who could help them.

C2 *Testimonial injustice*: subjects who complain are not believed by the officials, because they are part of a specific group. Or, the officials who treat the complaint demand documented evidence, of a kind that is not available for this group (e.g. proof of income). Usually, subjects are expert on their own situation. So here they are wronged in their capacity of knowing the relevant facts that matter to the case.

C3 *Hermeneutical injustice*: subjects who complain about a system error, know very well that something is wrong. But they are not experts in tax law and do not know the formal legal terminology (e.g. tax debt; evidence) in terms of which to analyze their own situation and formulate a complaint.

C4 *Content-based injustice*: subjects who file a complaint about the system, thereby claim that the system is wrong. The people responsible for the system, believe the system cannot be wrong. After all, "the system has been carefully developed, and has been tested by experts, etc" The complaint is much harder to understand, than the alternative (cognitive dissonance). Moreover, the complaint contradicts the consensus opinion among the system experts. People will not actively seek evidence to disproof that consensus (confirmation bias). That means, that the complaint must be wrong or even insincere.

These types of epistemic injustice can analyzed in a general model of knowledge and information exchange. We will now discuss a logic, to provide such a model.

3. Epistemic Injustice: An Epistemic Logic Perspective

Epistemic injustice goes beyond the wrongful recognition of an individual's epistemic status, it also examines how this misrecognition can lead to unfairness. Here we conceptualize that as unsatisfactory decisions for a person making a request for action. In this section, we will explore both the epistemic and action-oriented elements of epistemic injustice by introducing an action-based epistemic logic.

In this logic, the language we use to address the types of epistemic injustice includes *individual* knowledge K_a, *individual* belief B_a, and *common* belief C_G. We also include the modality E_a to illustrate *actions* or decisions.

Definition 1 (Language)**.** Let $Prop$ be a countable set of atomic propositions, and $\mathscr{I}$ be a finite set of agents. The language $\mathscr{L}$ is defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid B_a\varphi \mid C_G\varphi \mid E_a\varphi \mid \Box\varphi,$$

where $p \in Prop$, $a \in \mathscr{I}$, and $\emptyset \neq G \subseteq \mathscr{I}$.

The dual of K_a, denoted as $\hat{K}_a$, illustrates the consistency with agent a's knowledge. So $\hat{K}_a\varphi$ is defined as $\neg K_a\neg\varphi$ and read as "φ is consistent with agent a's knowledge." Similarly, the dual of belief $\hat{B}_a\varphi$ is defined as $\neg B_a\neg\varphi$ and read as "φ is consistent with agent a's belief," and the dual of universal modality $\Box$ is the existential modality $\Diamond$ such that $\Diamond\varphi := \neg\Box\neg\varphi$. The B-modality is a $\mathsf{KD4}$-modality, K- and $\Box$-modalities are $\mathsf{S5}$-modalities, and C_G-modality is a $\mathsf{KD4}$-modality.

The request by the sender or the applicant is dealt with by the decision maker. The decision maker has the right of decision to decide whether to fulfill the request or not. To express the fulfillment of the right of decision, our language includes an additional operator, denoted as E_b, to represent agent b's action execution. Thus, $E_b\psi$ can be interpreted as "Agent b takes specific actions to ensure that ψ holds true," or, from the perspective of agency theory or STIT logic [8, 4], "Agent b ensures that ψ is the case." In this context, we adopt Chellas's proposal [7] and treat this action operator, E_b, simply as a T-operator. The dual of E_b is denoted as $\hat{E}_b$, and $\hat{E}_b\psi$ is equal to $\neg E_b\neg\psi$.

Definition 2 (Models)**.** A structure $M = (W, \{R_a\}_{a\in\mathscr{I}}, \{D_a\}_{a\in\mathscr{I}}, \{\sim_a\}_{a\in\mathscr{I}}, V)$ is a model when it satisfies the following conditions:

- W is a non-empty set of states;
- R_a is an equivalence relation over W;
- D_a is a transitive and serial relation over W, such that $D_a \subseteq R_a$;
- $\sim_a$ is an equivalence relation over W;
- $V : Prop \to \wp(W)$ is a valuation function.

The accessibility relation R_a interprets individual a's knowledge, D_a represents individual a's belief, and $\sim_a$ represents individual a's ability to execute actions. We can define the transitive closure of all individuals' beliefs in the group G as $D_G = (\bigcup_{a\in\mathscr{I}} D_a)^+$. Now our modalities can be interpreted as usual:

$$M,w \models K_a\varphi \text{ iff } R_a[w] \subseteq ||\varphi||, \qquad M,w \models E_a\varphi \text{ iff } \sim_a [w] \subseteq ||\varphi||,$$
$$M,w \models B_a\varphi \text{ iff } D_a[w] \subseteq ||\varphi||, \qquad M,w \models \Box\varphi \text{ iff } W \subseteq ||\varphi||,$$
$$M,w \models C_G\varphi \text{ iff } D_G[w] \subseteq ||\varphi||,$$

where $||\varphi|| = \{w \in W \mid M,w \models \varphi\}$. So the following statements are valid:

CB $\quad C_G\varphi \to B_a\varphi$ if $a \in G$; $\qquad\qquad$ KB $\quad K_a\varphi \to B_a\varphi$.

By applying this action-based epistemic logic, we are able to define, for instance, the epistemic aspects of Fricker's notion of testimonial injustice, which involve the incorrect recognition of one's capacity as a knower.

- Agent b wrongly recognizes agent a's knowledge: $K_a\varphi \wedge B_b\neg K_a\varphi$;
- Agent b wrongly recognizes agent a's credibility of knowledge: $K_a\varphi \wedge \neg B_b K_a\varphi$.

In the next section, we will explore the formalization of the four types of epistemic injustice within our epistemic logic.

4. Towards a Formal Theory of Epistemic Injustice

We present four assumptions, (A1) – (A4), that characterize the situation of a decision maker b, who, as recipient of a statement φ, doubts the credibility of the sender a and subsequently rejects b's request, especially when sender $a \in A$ is perceived to be outside the privileged group B, and $b \in B$ (see Figure 1). We introduce notation $A \leqslant B$ to indicate that group A holds a disadvantaged epistemic position relative to group B. In this paper, we have not introduced a semantics for the expressions such as $A \leqslant B$. However, it's

worth noting that this is a viable possibility[3] [4]. The framework conditions provided here offer valuable insights for undertaking this task.

Within this context, $a \in A$ represents the message sender, while $b \in B$ signifies the message receiver and decision maker. In addition, φ represents the evidence submitted by a and ψ represents the requested decision to be made by b. Proposition $\Box(\psi \to \varphi)$ represents that "Evidence φ submitted by a leads to fulfillment of a request for ψ". This means that the evidence is a necessary condition for fulfilling the request. The formalization of the four assumptions are as follows:[5]

A1 For all $b \in B$: $\Box(\psi \to \varphi) \to (E_b\psi \to B_b\varphi)$;
A2 For all $a \in A$ and $b \in B$ with $A \leqslant B$: $B_b\varphi \to K_a\varphi$;
A3 For all $a \in A$ and $b \in B$ with $A \leqslant B$: $B_b\varphi \to B_bK_a\varphi$;
A4 For all $b \in B$: $C_B\neg\varphi \to B_b\neg\hat{E}_bC_B\varphi$.

These four assumptions serve to elucidate the underlying factors leading to *prejudice* against individuals in disadvantaged positions by those in privileged positions. Prejudice, as delineated by these four assumptions, is not solely a manifestation of power owned by privileged decision-makers. It also arises from the presence of several *irrational* assumptions underlying their decision-making processes. These irrationalities become apparent in the assumptions we have outlined above. The clarifications of the above assumptions are given as follows:

Assumption (A1) captures the decision rights of b to grant ψ. Here $E_b\psi \to B_b\varphi$ means that b believing submitted evidence φ is a necessary condition for b to ensure ψ.

Assumption (A2) reflects the irrationality of the advantageous and dominant position held by the receiver: If the decision maker $b \in B$, positioned in an advantaged state B (which is illustrated as $A \leqslant B$ [6]), and holds a certain belief, it serves as a compelling rationale to posit that members of the disadvantaged group must possess the same knowledge. In essence, the beliefs of the privileged party take precedence over the knowledge of the disadvantaged party. Further, the concept of *prejudice* is exemplified by this interdependence: When an individual lacks knowledge of a certain aspect φ, this becomes a reason that the information sent by this agent is not believed by the decision maker, primarily due to their skepticism toward the disadvantaged group (i.e. $A \leqslant B$).

[3] Intuitively, $A \leqslant B$ when according to any $b \in B$, any $a \in A$ believes fewer formulas than $b \in B$. Let $\mathsf{Form}_w(b,a) = \{\varphi \mid M, w \models B_bK_a\varphi\}$. We have $M, w \models A \leqslant B$ iff $\mathsf{Form}_w(b,a) \subseteq \mathsf{Form}_w(b,b)$ for all $a \in A, b \in B$.

[4] In this setting, notation $A \leqslant B$ is relative to the topic area of proposition φ. Suppose *Prop* is divided in overlapping subsets $T \subset Prop$, that denote a topic area, such as finance, or sports, etc. A formula can be classified by the topic of the proposition letters in it. Now in general, a person b trusts a person a to know φ whenever the topic of φ is in the competence areas of person a, according to be b. See [11]

[5] The frame conditions to validate (A1) – (A4) are as follows:

(A1) $\forall wu \in W(wD_bu \to w \sim_b u)$;
(A2) $\forall wu \in W(wD_bu \to wR_au)$, if $A \leqslant B$, $a \in A$ and $b \in B$;
(A3) $\forall wuv \in W(wD_bu \wedge uR_av \to wD_bv)$, if $A \leqslant B$, $a \in A$ and $b \in B$;
(A4) $\forall wuv \in W(wD_bu \wedge u \sim_b v \to wD_bv)$.

[6] In this paper, we introduce a binary relation denoted as $\leqslant$ to illustrate intergroup *prejudice*, providing a simplified representation for this key concept in epistemic injustice. While it's acknowledged that prejudice in the real world can be influenced additionally by various factors, such as topics that are discussed in [11], our current focus centers on establishing the logical principles for defining prejudice between groups. The examination of prejudice with respect to both groups and topics remains a subject for future research.

Assumption (A3) highlights the dominant position of the advantaged group from a different perspective. When an individual within the advantaged group accepts a piece of information, they believe that any member in the disadvantaged group must possess this information as their knowledge. This phenomenon underscores the concept of "*prejudice*" as one type of interdependence of communication: The beliefs of the dominant individual influence their perceptions regarding the knowledge of those in the disadvantaged group.

Assumption (A4) sheds light on the concept of common ground within the privileged group. Simply speaking, when a piece of information is established as part of the common ground for the group, every individual within that group believes it is impossible to revise such a common belief.

To understand epistemic injustice, these four assumptions play a pivotal role, as outlined in Table 2. Note that the reasoning behind distributional injustice, testimonial injustice, and content-based injustice differs in three key aspects, respectively: distinctions in factual information, beliefs of decision-makers about credibility of groups, and beliefs of decision-makers about credibility of content.

Distributional injustice is rooted in the fact that the sender a, who is in a disadvantaged position, lacks knowledge of the evidence φ, which can be expressed as $\neg K_a\varphi$. Given this fact and the communication assumption (A2), it leads to "*weak belief*": the decision maker b does not believe in the evidence φ submitted by agent a, denoted as $\neg B_b\varphi$. This type of belief is considered *weak*, because it is derived from a basis of the other's lack of knowledge [26, 21]. Following assumption (A1), which addresses the decision right of b, the decision doesn't fulfill the request ψ from agent a: $\neg E_b\psi$.

In contrast, the reasoning process for testimonial injustice follows a different path. While it also involves a weak belief $\neg B_b\varphi$, it is inferred from a distinct basis of information. Testimonial injustice is rooted in the fact that the sender a indeed possesses knowledge of the evidence φ (i.e., $K_a\varphi$). It also relies on the epistemology assumption that the decision-maker does believe the sender genuinely lacks knowledge of the evidence, denoted as $B_b\neg K_a\varphi$. This belief is referred to as "*strong belief*," because it is assumed and not derived. This strong belief is labeled as a *prejudice*, because it presupposes that **everyone** in a disadvantaged group A lacks knowledge about this topic area, regardless of its actual veracity. From this strong belief, assumption A2 and axiom D, the weak belief $\neg B_b\varphi$ can also be inferred. Ultimately, assumption A3 leads to non-fulfillment of the request.

Concerning the last row of Table 2, content-based injustice, we can model two cases: (i) the individual case $B_b\neg\varphi$ and therefore $\neg B_b\varphi$, so the request is rejected, and also $B_b\neg K_a\varphi$, so the requester is denied in her right as a knower, and (ii) the group consensus case, $C_B\neg\varphi$ and therefore $\neg C_B\varphi$, so the request would be rejected by any official, but also $C_B\neg K_a\varphi$, so the requester is by consensus denied in the right as a knower.

Note however, that in this simple form of epistemic logic we cannot distinguish between the failure of b to understand φ (cognitive dissonance), and failure of b to publicly accept statement φ as true , in the group (confirmation bias). We cannot express the need for a belief revision either, if φ would be accepted as true in case $\neg\varphi$ is already believed, which would take effort. For a similar reason, we do not even attempt to capture (C3) hermeneutical injustice. In our logic, we cannot express that only some agents have an ontology to understand (for b) or to express (for a) a problem situation.

	Facts	Beliefs	Request Fulfillment	Inferential Elements
Distributive Injustice	$\Box(\psi \to \varphi)$ $\neg K_a \Box(\psi \to \varphi)$ $\neg K_a \varphi$	Weak: $\neg B_b \varphi$	 $\neg E_b \psi$	 A2, $\neg K_a \varphi$ A1, $\Box(\psi \to \varphi)$, Weak
Testimonial Injustice	$\Box(\psi \to \varphi)$ $K_a \Box(\psi \to \varphi)$ $K_a \varphi$	Strong: $B_b \neg K_a \varphi$ Weak: $\neg B_b \varphi$	 $\neg E_b \psi$	 A3, D, Strong A1, Weak
Content-based Injustice	$\Box(\psi \to \varphi)$ $K_a \Box(\psi \to \varphi)$ $K_a \varphi$	Strong: $C_B \neg \varphi$ Weak$_1$: $C_B \neg K_a \varphi$ Weak$_2$: $B_b \neg K_a \varphi$ Weak$_3$: $\neg B_b \varphi$ Weak$_4$: $B_b \neg \hat{E}_b C_B \varphi$	 $\neg E_b \psi$	 T, NEC$_C$, Strong CB, Weak$_1$ A3, D, Weak$_2$ A1, Weak$_3$ A4, Strong

Table 2. A Classification of epistemic injustice (C1,C2,C4), where $a \in A$ and $b \in B$ with $A \leqslant B$.

There is another kind of case when b would be inclined to believe that φ based on a statement by a, but still believes that it would be impossible to convince the others: $C_B \neg \varphi$ but $B_b \Diamond \varphi$ (while, let's say, most of the group members also believe that it is actually $\Box \neg \varphi$), so also $B_b \Diamond K_a \varphi$, but still $B_b \neg \Diamond C_B \varphi$ (or just $B_b \neg \Diamond E_b C_B \varphi$) so he doesn't do anything.

5. Cases

Case 1. Toeslagenaffaire (Netherlands) The Toeslagenaffaire (child care benefits scandal) is a complex set of interrelated cases and problems, of a political, legal, technical and administrative nature, in the Netherlands in the period 2010-2017 [10], although the consequences still haven't been fully dealt with. This is a complex and sensitive case. Here we can only provide a few telling observations. Observations are **bold** in the text.

The agency executing the child care benefit scheme, is the Netherlands Tax Administration, specifically the department Benefits (Belastingdienst/Toeslagen). At the time, the tax office was seen as more competent in IT, than other government agencies.

The benefit scheme started with the wish of politicians to create a market for child care and to stimulate women to get paid work. The state funds child care centres indirectly, by reimbursing those parents for the costs incurred, whose income is below a certain threshold. This may involve several hundreds of Euros per month. To get reimbursed, parents have to apply for child care benefit. Legally, child care benefit is a conditional entitlement. Parents are only entitled to a certain amount of the benefit, if they actually use childcare for a certain number of hours, if the care centre is approved, and if their combined income stays below a certain threshold. To provide evidence of these conditions, parents have to fill out forms and supply evidential documents, often obtained from other parties, like the care centre (number of hours). Given the complexity of the forms and rules, it is likely that mistakes are made. Moreover, many people do not know in advance exactly how much income they will earn. Social benefit agencies are used to

working with such estimates. However, for Toeslagen, the burden of proof was put on the families: by law, they are responsible for providing exact numbers about their situation.

An important factor that drove this attitude was the political pressure to combat fraud, in the years leading up to the scheme. Here we highlight the so-called black list. The tax office generally applies risk-based supervision [5]. However, here **subjective risk indicators** were used. For example, a person owns an expensive car without the income to support it. Indicators were not verified to be effective for finding fraud. Some risk indicators, like nationality, were later ruled by the AP (Data Protection Agency) to be ineffective, unnecessary for the task and therefore unlawful (GDPR; WBP) [3]. The use of nationality was also ruled to be discriminatory [3]. Originally, the black list was used internally by fraud teams, as a warning to colleagues to look into. Later, from 2014, the list was also used in Benefits for regular application processing. Being on the blacklist itself became reason to be denied child care benefit. Citizens received **no explanation** for such rejections. Naturally, these citizens complained, but many of these **complaints were ignored** or rejected. This shows a pattern [1]:

> "In the CAF 11 case, the focus on fighting fraud caused **institutional bias**, according to the committee. That bias meant that from the outset, the actions of Benefits were based on the suspicion that the CAF 11 parents had committed fraud. A suspicion that was not based on the personal actions of these parents, but on the mere fact that they were being monitored as part of the CAF 11 file." Translation of [1].

Here 'CAF 11 file' refers to the black list. CAF refers to the name of the fraud team.

The term *institutional bias* (NL: institutionele vooringenomenheid), from the parliamentary committee report, created a lot of political discussion. It means that Benefits was prejudiced against a group of parents, specifically those listed on the CAF 11 file. The bias was institutional, because it was given a place in work instructions and also worked through in objection and appeal procedures, in the recovery measures as well as in new applications for childcare benefits from these parents. For instance, hearings revealed that management overseeing the CAF team accepted that, if about 80% of the people on the list were indeed 'bad guys', 20% of people on the list were 'good guys', and thus unjustifiably targeted (80/20 rule). ([10, p 47].

The committee describes that after terminating a benefit, the tax office would investigate in detail, whether parents had been entitled to the benefits received. This investigation was deliberately **aimed at shortcomings** - even the slightest ones - in administration, payments or supporting documents, with the purpose to have the benefit withdrawn. Citizens normally get an opportunity to correct their statements, except when the tax office believes manipulations are intentional. Here, citizens were given **no opportunity to correct**. So in effect all citizens making mistakes were treated as criminals.

If an administrative discrepancy was discovered, the entire amount of benefits had to be paid back, not just a correction for the amount involved (**all or nothing approach**). This increased the impact for families. Later, legal scholars agreed this was disproportional [10]. Moreover, to many families this felt like a punishment.

The text of the law, does not give officials discretionary power to deviate from the policies. For instance, there is no hardship clause, as customary in social benefit. The strict application of the law by Toeslagen, was also confirmed in case law. The Council of State (Raad van State) often ruled in favour of Toeslagen, in appeal cases in which the strict application of the right to childcare benefits was questioned. Internal doubt in the

Observation	Effect	Type of epistemic injustice
subjective risk indicators	decision based on group, not evidence	distributional
no explanation	citizens no knowledge what went wrong	distributional
complaints ignored	officials no knowledge of problems	content-based, testimonial
institutional bias	decisions based on systematic prejudice	distributional; testimonial
aimed at shortcomings	mistakes treated as violations	hermeneutical
no opportunity to correct	mistakes treated as criminal	testimonial; content-based
all-or-nothing	disproportional impact of violation	hermeneutical
no redress mechanism	increased duration, impact of harm	hermeneutical; testimonial

Table 3. Case observations analyzed as epistemic injustice

tax office about the hardship, were silenced, with reference to the ruling by the Council of State. So initially, there were **no mechanisms for redress or appeal**. This greatly extended the duration and impact of the hardships sustained by families.

Analysis Table 3 repeats the main observations and shows that they constitute a case of epistemic injustice. To summarize, all forms of epistemic injustice were present in this case. Distributional and testimonial are most common, but also cases of content-based and hermeneutical were found, especially in the aftermath and the inability of the administration to handle complaints and redress the problems.

5.1. Case 2. Robodebt (Australia)

For lack of space, we cannot make a similarly detailed analysis of the Centrelink case in Australia (New South Wales), that became known as 'RoboDebt' [19]. The Centrelink system was supposed to automatically calculate the tax debt of a citizen, based on evidence provided by that citizen and on files already in possession of the tax office. However, in case the financial figures were absent, the system used heuristics and machine learning to estimate the missing data. Although unverified, these estimates were used to calculate the tax debt, and consequently, the taxes to be paid. Some people had to pay a large amount of taxes they didn't owe. Victims found it impossible to understand what they had done wrong, and what caused these large debts. Some victims stated they thought they had lost their mind: it was their word against that of the tax office.

Clearly, there are many similarities between the two cases. Both were concerned with the tax office and a strict interpretation of the law, concerning financial evidence. Both projects were initiated under political pressure to get results, for which the computer system wasn't originally intended. Both show a large trust in technology from the government, and a lack of understanding for the situation of citizens. In both cases it took a long time for victims to be heard and to get compensation. There are also interesting differences. The Dutch case focused on fraud detection in child care benefits, whereas the Australian case is about tax debt, so a different legal framework.

6. Conclusions and future work

In this paper we have taken a complex and sensitive topic from the field of social epistemology, namely epistemic injustice.

We have defined epistemic justice and identified four types, based on the literature [15, 16, 14]: distributional, testimonial, hermeneutical an content-based epistemic injus-

tice. We have analyzed the mechanisms behind these types, using speech act theory, and provide a formal characterization of three of these types in an action-based epistemic logic for beliefs and knowledge, namely distributional, testimonial and content-based injustice. We present four assumptions as foundational elements in the development of a formal theory encompassing these three types of epistemic injustice. These assumptions elucidate the inferential components at play in the reasoning processes of privileged individuals.

The transformation into epistemic logic is certainly not finished. For example, so far, we cannot express the hermeneutical type, because we have no notation for conceptual analysis and logical expression. In addition, we cannot properly express the notion of social power, nor the notion of identity prejudice, which plays a crucial role in Fricker's original work. Currently, this part of the analysis is based on a simplified setting (Figure 1). In the future, we would like to generalize and add a notation for background facts about roles and social relations between agents, that influence epistemic trust and prejudice.

We have analyzed one case about a government decision making system [10], and list observations that correspond to each of the four types of epistemic injustice. In future work, we want to analyze the second case [19] in more detail.

We welcome further debate about the methodology of the paper and hope that it will also invite others to broaden the scope of topics for logical analysis.

References

[1] Adviescommissie Uitvoering Toeslagen. Omzien in verwondering. Tweede Kamer der Staten Generaal, 2019.

[2] J. L. Austin. *How to do things with words*. Harvard UP, Cambridge MA, 1962.

[3] Autoriteit Persoonsgegevens. Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag. Report z2018-22445, 2020.

[4] Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, 2001.

[5] Julia Black. The emergence of risk-based regulation and the new public risk management in the united kingdom, 512–48. *Public Law*, Autumn:512–548, 2005.

[6] H. Carel and I.J. Kidd. Epistemic injustice in healthcare: a philosophical analysis. *Medicine, Health Care and Philosophy*, 17:529–540, 2014.

[7] Brian F Chellas. *Modal logic: an introduction*. Cambridge university press, 1980.

[8] Brian Farrell Chellas. *The logical form of imperatives*. Stanford University, 1969.

[9] Herbert H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

[10] C.J.L. van Dam. Ongekend onrecht - verslag van de parlementaire ondervraging kinderopvangtoeslag. Tweede Kamer der Staten Generaal, 2020.

[11] M. Dastani, A. Herzig, J. Hulstijn, and L. van der Torre. Inferring trust. In J. Leite, editor, *Proceedings of Fifth Workshop on Computational Logic in Multi-agent Systems (CLIMA V)*, pages 144–160. Springer-Verlag, Berlin, 2004.

[12] B. De Bruin. Epistemic virtues in business. *Journal of Business Ethics*, 113:583–595, 2013.

[13] Boudewijn de Bruin. Epistemic injustice in finance. *Topoi*, 40:755–763, 2021.

[14] Robin Dembroff and Dennis Whitcomb. Content-focused epistemic injustice. In Tamar Szabó Gendler, John Hawthorne, and Julianne Chung, editors, *Oxford Studies in Epistemology - Volume 7*. Oxford University Press, 2022.

[15] Miranda Fricker. *Epistemic injustice.* Oxford University Press, Oxford, 2007.

[16] Miranda Fricker. Epistemic justice as a condition of political freedom? *Synthese*, 190:1317–1332, 2013.

[17] Alvin Goldman and Cailin O'Connor. Social epistemology. *Stanford Encyclopedia of Philosophy*, 2021.

[18] Ben Green and Amba Kak. The false comfort of human oversight as an antidote to A.I. harm. *Slate*, June, 2021.

[19] Catherine Holmes. Report of the royal commission into the robodebt scheme. Report, Commonwealth of Australia, 2023.

[20] Christopher Hookway. Some varieties of epistemic injustice: Reflections on fricker. *Episteme*, 7:151–163, 2010.

[21] Thomas Icard, Eric Pacuit, and Yoav Shoham. Joint revision of beliefs and intention. In *KR*, 2010.

[22] Ian James Kidd, José Medina, and Gaile Pohlhaus Jr., editors. *Handbook of Epistemic Injustice*. Routledge, London, 1 edition, 2017.

[23] Ulrik B.U. Roehl. Automated decision-making and good administration. *Government Information Quarterly*, 40(4):101864, 2023.

[24] John R. Searle. *Speech acts: an Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

[25] Haroon Sheikh, Corien Prins, and Erik Schrijvers. *Mission AI*. Netherlands Council for Government Policy (WRR). Springer, The Hague, 2023.

[26] Wiebe van Der Hoek, Wojciech Jamroga, and Michael Wooldridge. Towards a theory of intention revision. *Synthese*, 155:265–290, 2007.