

On Committing Authenticated-Encryption

John Chan and Phillip Rogaway

Department of Computer Science
University of California, Davis, USA

Abstract. We provide a strong definition for *committing* authenticated-encryption (cAE), as well as a framework that encompasses earlier and weaker definitions. The framework attends not only to *what* is committed but also the extent to which the adversary knows or controls keys. We slot into our framework strengthened cAE-attacks on GCM and OCB. Our main result is a simple and efficient construction, CTX, that makes a nonce-based AE (nAE) scheme committing. The transformed scheme achieves the strongest security notion in our framework. Just the same, the added computational cost (on top of the nAE scheme’s cost) is a single hash over a short string, a cost independent of the plaintext’s length. And there is *no* increase in ciphertext length compared to the base nAE scheme. That such a thing is possible, let alone easy, upends the (incorrect) intuition that you can’t commit to a plaintext or ciphertext without hashing one or the other. And it motivates a simple and practical tweak to AE-schemes to make them committing.

Keywords: AEAD, authenticated encryption, committing encryption, key-robustness

1 Introduction

A natural misconception about authenticated encryption (AE) is the belief that a ciphertext produced by encrypting a plaintext with a key, nonce, and associated data (AD) effectively *commits* to those things: decrypting it with some *other* key, nonce, or AD will usually fail, the transmission deemed invalid. And why not? One wouldn’t expect to successfully open a lock when using an incorrect key. The intuition is even memorialized in the name *authenticated* encryption: things aren’t just private, the name implies, but authentic.

Yet Farshim, Orlandi, and Rošie [10] (FOR17) point out that AE provides no such guarantee—not if the adversary can select any keys. Subsequent work demonstrated that just *knowing* the keys suffices to construct a ciphertext that decrypts into different valid messages [8,12]. A variety of work has also made clear just how wrong things can go when designers implicitly and incorrectly assume that their encryption *is* committing [2,8,14].

We call the event of a ciphertext being “explained” in multiple and valid ways a *misattribution*. The cited works offer definitions and schemes that seek to protect against misattribution. But these definitions are mostly incomparable, weak, and fold in aims beyond avoiding misattribution.

DEFINITIONAL FRAMEWORK. To begin, we revisit definitions for committing AE. We offer a definitional framework that unifies and strengthens previous definitions targeting misattribution. We call the security goals *committing-AE* (cAE). The framework applies to schemes for *nonce-based AE with associated data* (nAE). Encryption takes in a key K , a nonce N , an associated data A , and a message M , and outputs a ciphertext C . Under our framework, an adversary succeeds in an attack when it creates a misattribution. That happens when C results from known and distinct tuples (K, N, A, M) and (K', N', A', M') for valid messages M, M' . We say “results from” because C could be output by encryption *or* input to decryption—anything that results in adversarial knowledge of the pair $(K, N, A, M), (K', N', A', M')$.

Previous definitions consider only some forms of misattribution. For example, the *full robustness* and *key-commitment* notions [2,10] require that the keys differ, $K \neq K'$, but ignore the possibility of misattribution under the same key. Our framework can encompass all possible types of misattribution (see Appendix A). That said, we regard the desired target as the *strongest* definition, AE that is *fully committing*, where the adversary wins if it manages *any* form of misattribution.

Our framework attends also to the status of keys held by parties. To model different levels of adversarial activity, we include a definitional parameter \mathbf{t} . This two-character string dictates what types of keys the adversary might employ for a misattribution to occur. Keys are either: *honest* (represented by the character 0), meaning they are generated uniformly at random and remain unknown to the adversary; *revealed* (represented by a 1), meaning they were honestly generated, but the adversary knows their value; or *corrupted* (an X), meaning the adversary itself chose the key. This gives rise to six different definitions. This “knob” is useful for describing and understanding attacks. The weakest of these notions models when both keys are honest. We show that ordinary nAE-security implies this notion assuming the adversaries do not repeat nonces for the same key. For some applications that require cAE security, notions weaker than the strongest notion (XX-security) may work just fine. In such cases, one might be able to obtain stronger quantitative bounds.

MAIN CONSTRUCTION. Our main result is a method to convert an arbitrary (tag-based) nAE scheme into a similarly efficient cAE scheme. We set high bars for security and efficiency. Security is with respect to the strongest form of commitment: K, N, A , and M must all be “fixed” by a ciphertext, even if the adversary controls all keys.

Our CTX construction is extremely simple. Starting from an nAE scheme whose encryption algorithm $\mathcal{E}(K, N, A, M)$ produces a ciphertext $C = C \parallel T$ consisting of a ciphertext core C (with $|C| = |M|$) and a tag T (with $|T| = \tau$), just replace the tag T with an alternative tag $T^* = H(K, N, A, T)$ (this tag of length μ). Decryption does the obvious, verifying T^* . The function H is a cryptographic hash function that, in the security proofs, is modeled as a random oracle. The remarkable fact is that this extremely simple tweak to the nAE scheme not only works to commit to K, N , and A , but also to the underlying

message M . This ultimately follows from the injectivity of the map from the ciphertext core C to the plaintext M when K , N , and A are all fixed.

The CTX construction is computationally efficient insofar as the work on top of the base nAE scheme is a hash computation over a string that does not grow with the plaintext or ciphertext. And the nAE scheme’s minimal ciphertext expansion is preserved, going from the τ (typically 128) extra bits that are needed to provide authenticity to the μ (typically 160) extra bits that are needed to provide authenticity *and* the binding (commitment) of all inputs.

ATTACKS ON GCM AND OCB. Previous misattribution attacks on GCM were mounted with adversarial control of the keys [8,12]. It is mentioned by those same authors that knowledge of the keys is sufficient. Under our terminology, this would be a CAE_{xx}-attack and a CAE₁₁-attack respectively.

We present a new attack on GCM for a weaker adversary, a CAE₀₁-attack. That is, the adversary can create a misattribution knowing just one key. For any ciphertext C generated under a perfectly honest key, one can find a valid decryption for it under a known key. The attack strategy involves computing an AD that validates the decryption of the ciphertext. Intuitively, for any key, nonce, message, and ciphertext, there are an infinite number of ADs that validly decrypt the ciphertext—we only need to find one of them. The strategy extends to mounting a CAE₀₁-attack on OCB as well. These attacks demonstrate that nAE-security is insufficient for even CAE₀₁-security.

RELATED WORK. Prior work has been leading towards a definition for fully committing AE (the cAE-xx notion), but didn’t quite get there. There has also been movement towards efficient schemes for this end.

The notion of committing encryption goes back to 2003 with Gertner and Herzberg [11], who consider the problem in both the symmetric and asymmetric settings. The authors do not look at deterministic or authenticated encryption.

Abdalla, Bellare, and Neven give definitions for what they term *robustness* [1]. The work is in the asymmetric setting and requires an adversary to produce a ciphertext that validly decrypts under two different keys. Their notion encompasses keys that are honestly generated. Later, Farshim, Libert, Paterson, and Quaglia point out that, for some applications, robustness against adversarially-chosen keys is critical [9]. They strengthen Abdalla et al.’s notion to address this observation.

Farshim, Orlandi, and Rosié (FOR17) [10] contextualized Abdalla et al.’s robustness in the AE setting, initializing the study of what we call committing AE. Shortly after, Grubbs, Lu, and Ristenpart (GLR17) [12] defined a variant of committing AE with the goal of constructing schemes that support *message franking*. Dodis, Grubbs, Ristenpart, and Woodage (DGRW18) [8] also target message franking and further develop GLR17’s definitions. These two works have goals beyond preventing misattributions. We are after simpler aims, with the syntax of classical nAE. Albertini, Duong, Gueron, Kölbl, Luykx, and Schmieg (ADGKLS20) [2] observe the possibility of mitigating the attacks described by GLR17 and DGRW18 under a weaker form of misattribution prevention. Their

observation led them to develop a more efficient construction—one that avoids additional passes over the message.

Bellare and Hoang (BH22), in a contemporary work, offer a range of committing AE definitions, with starting points of both standard nAE and misuse-resistant AE [5]. The strongest of their definitions, like ours, requires that the ciphertext commit to *everything*—the key, nonce, AD, and plaintext. They also consider *multi-input committing security*, where an adversary is required to create misattributions of more than just two valid explanations.

Len, Grubbs, and Ristenpart demonstrate password-recovery attacks on non-committing password-based AEAD schemes [14]. Their attacks are built on efficiently creating ciphertexts that successfully decrypt under many keys.

A more detailed comparison of some of the cited works is in Section 6.

2 Preliminaries

COLLISION-RESISTANT HASH FUNCTIONS. A *hash function* $H: \mathcal{D} \rightarrow \{0,1\}^h$ maps strings from some domain $\mathcal{D} \subset \{0,1\}^*$ to strings of length h . Informally, a hash function is collision-resistant if it is difficult for an adversary \mathcal{A} to find two unique inputs that map to the same output. This notion is captured by a collision resistance game CR where \mathcal{A} is ran and outputs a pair (M, M') . The game outputs **true** if $H(M) = H(M')$ and $M \neq M'$. The adversary \mathcal{A} 's advantage against H is then quantified as $\mathbf{Adv}_H^{\text{col}}(\mathcal{A}) = \Pr[\text{CR}_H^{\mathcal{A}} \Rightarrow \text{true}]$. This definition of collision-resistance of unkeyed hash functions follows the human-ignorance approach of [18].

NONCE-BASED AE. An nAE scheme, or a *nonce-based authenticated encryption-scheme supporting associated data (AD)* is a pair of functions $(\mathcal{E}, \mathcal{D})$. The former, the *encryption algorithm*, is a deterministic function $\mathcal{E}: \mathcal{K} \times \mathcal{N} \times \mathcal{A} \times \mathcal{M} \rightarrow \mathcal{C}$ that takes in a key, a nonce, an AD, and a message, and outputs a ciphertext. The latter \mathcal{D} , the *decryption algorithm*, is a deterministic function $\mathcal{D}: \mathcal{K} \times \mathcal{N} \times \mathcal{A} \times \mathcal{C} \rightarrow \mathcal{M} \times \{\perp\}$. We sometimes write $\mathcal{E}_K^{N,A}(M)$ and $\mathcal{D}_K^{N,A}(C)$ to denote $\mathcal{E}(K, N, A, M)$ and $\mathcal{D}(K, N, A, C)$. An nAE scheme is *correct* if $\mathcal{D}_K^{N,A}(\mathcal{E}_K^{N,A}(M)) = M$ for all K, N, A, M . A notable property of correct schemes is how encryption is injective from \mathcal{M} to \mathcal{C} when K, N, A are fixed. $K \in \mathcal{K}, N \in \mathcal{N}, A \in \mathcal{A}, M \in \mathcal{M}$ and $\mathcal{D}_K^{N,A}(C) = \perp$ otherwise. We assume that the message space $\mathcal{M} \subseteq \{0,1\}^*$ is a set of strings where $M \in \mathcal{M}$ implies $\{0,1\}^{|M|} \in \mathcal{M}$. We insist of an nAE scheme that there is an associated value, its *expansion*, which is a constant τ such that $|\mathcal{E}_K^{N,A}(M)| = |M| + \tau$ for all K, N, A, M .

A customary formulation of nAE security asks an adversary attacking the nAE scheme $\Pi = (\mathcal{E}, \mathcal{D})$ to distinguish between a pair of oracles [17]. The “real” oracles use Π 's algorithms while the “ideal” or “fake” oracles only give bogus responses. For an adversary \mathcal{A} attacking Π , its advantage is defined as follows:

$$\begin{aligned} \mathbf{Adv}_{\Pi}^{\text{nae}}(\mathcal{A}) &= \Pr[K \leftarrow \mathcal{K}; \mathcal{A}^{E_K(\cdot,\cdot), D_K(\cdot,\cdot)} \Rightarrow 1] - \\ &\quad \Pr[\mathcal{A}^{\$, \perp} \Rightarrow 1]. \end{aligned}$$

Its interaction with the E_K and D_k oracles, the “real” oracles, begins with the uniformly random sampling of a key K . An oracle query then $E_K(N, A, M)$ returns $\mathcal{E}_K(N, A, M)$, while an oracle query $D_K(N, A, C)$ returns $\mathcal{D}_K(N, A, C)$. In contrast, an “ideal” oracle query of $\$(N, A, M)$ returns a uniformly random string of length $|M| + \tau$, while $\perp(N, A, C)$ always returns \perp .

The adversary is forbidden from querying its decryption oracles with N, A, C if it acquired C from its encryption oracles using N, A as doing so would allow it to trivially win. Similarly, the definition demands that adversaries are *nonce-respecting*, meaning that they never repeat a nonce in its encryption queries.

We will find it useful to define a variant of nAE security that directly models multiple keys, which was first formalized in [7]. In this variant, an infinite number of keys are uniformly randomly generated for the real oracles at the initialization of the security game. Each oracle takes in an additional parameter, an index, that the adversary uses to specify which key to use for its query. Its advantage notion in this game is:

$$\mathbf{Adv}_{\Pi}^{\text{nAE*}}(\mathcal{A}) = \Pr[\mathbf{K} \leftarrow \mathcal{K}^\infty; \mathcal{A}^{E_{\mathbf{K}}(\cdot, \cdot, \cdot), D_{\mathbf{K}}(\cdot, \cdot, \cdot)} \Rightarrow 1] - \Pr[\mathcal{A}^{\$(\cdot, \cdot, \cdot), \perp(\cdot, \cdot, \cdot)} \Rightarrow 1].$$

Similarly, the adversary is restricted from querying (i, N, A, C) to its decryption oracle if C is the result of some (i, N, A, M) . The adversary is nonce-respecting in this case if it never repeats the same nonce for the same key when querying the encryption oracle.

3 Committing AE

COMMITTING AE. Informally, we call an nAE scheme a *committing AE scheme* (cAE) if it commits to any of the elements used to produce a ciphertext. We are primarily interested in cAE schemes that commit to all of these elements. By the definition of nAE in section 2, those elements would be the key, nonce, AD, and message. The CAE game that captures this property is presented in Fig. 1.

An adversary attacking the CAE-security of an nAE scheme Π aims to produce a ciphertext C that has two distinct valid “explanations.” That is, ciphertext C could decrypt to a messages M using (K_i, N, A) , or it could decrypt to a message M' using (K_j, N', A') such that $(K_i, N, A, M) \neq (K_j, N', A', M')$ and $M, M' \neq \perp$. When either of these occur, we say that the ciphertext is *misattributed*. We sometimes refer to C as the *colliding ciphertext* and the (K, N, A, M) associated to it as one of its attributions. In the game code, we write $S \overset{\cup}{\leftarrow} \{x\}$ as shorthand for $S \leftarrow S \cup \{x\}$, adding x to the set S .

The adversary initializes the game with the Initialize procedure, which generates an infinite number of uniformly random keys indexed by the natural numbers. Several sets are also initialized, one of which is the set S which keeps track of (K, N, A, M, C) tuples that constitute encryption and decryption queries and responses made by the adversary. The game terminates with the Finalize procedure, which checks the tuples of S in a pairwise fashion for an adversarial win.

That is, it searches for a pair of tuples where the ciphertexts are equivalent and that the explanations are distinct and valid. There is an additional condition checked that pertains to the function chk that we describe later.

There are four other game procedures surfaced to the adversary: ENC, DEC, REV, COR. These are the encryption, decryption, reveal, and corruption oracles. The first two oracles let the adversary use Π 's encryption and decryption algorithms using a key specified by an index i . Any ciphertext or message generated by the call to Π 's algorithms is stored alongside the queried K_i, N, A, M (or C) are stored in the set S . The reveal oracle allows the adversary to query an index i and learn the key K_i . For the corruption oracle, the adversary queries an index i and a key K and supplants K_i with K . Keys that are affected by these two oracles are added to the sets K_r and K_c respectively.

Note that ENC queries are restricted to be *nonce-respecting* for honest keys. That is, an adversary cannot repeat nonces for its encryption queries to an honest key K_i . This is reflected in the game code. The purpose of this is to prevent possibilities of an adversary learning an honest key through abuse of the nonce as this would otherwise blur the distinction between revealed and honest keys.

When the adversary yields a colliding ciphertext with two distinct valid explanations, there is one more condition to check before the adversary is considered to have won. That is, the chk function is ran on the keys of the explanations. This function is the *collision check* function and relies on a parameter of the CAE game, t , which we refer to as the *collision type*. For any key K_i in the game, the key can either be *corrupted*, *revealed*, or *honest*. A key is corrupted when it is added to the game through the corruption oracle COR and thus part of the set K_c . A key is revealed when the adversary learns of it through the reveal oracle REV and thus part of the set K_r . If the key is part of neither set, meaning it was chosen uniformly at random and unaffected by the adversary, then it is considered honest. Whether keys are corrupted, revealed, and honest are represented by X, 1, and 0 bits respectively. Six different types of collisions arise from these types of keys. The parameter t is a two-bit string that describes the kind of collision the adversary may win with.

Finally, the advantage of an adversary \mathcal{A} attacking the CAE-security of an nAE scheme Π in regards to a collision type t is quantified as $\mathbf{Adv}_{\Pi,t}^{\text{cae}}(\mathcal{A}) = \Pr[\text{CAE}_{\Pi,t}^{\mathcal{A}} \rightarrow 1]$. When discussing CAE-security with a specific type of collision, we denote the collision with a subscript i.e. CAE_{xx}-security.

OTHER COMMITTING NOTIONS. Most other committing AE definitions focused on cases where the adversary has control over both keys when creating a colliding ciphertext, which would be a corrupted-corrupted (or $t = \text{XX}$) collision [2,8,12]. Farshim et al. consider one definition of key-robustness, called *semi-full robustness*, where the adversary is asked to come up with a ciphertext that decrypts under an honest key and a key that it knows (what we would call a 01-collision) [10]. Bellare gives another robustness notion for randomized symmetric encryption called *random-key robustness* in [4] that is comparable to the CAE₀₀ notion and shows that authenticity implies random-key robustness. Our defini-

$\text{CAE}_{\Pi, \tau}$

procedure Initialize() 00 for $i \in \mathbb{N}$ do $K_i \leftarrow \mathcal{K}; N_i \leftarrow \emptyset$ 01 $S, K_c, K_r \leftarrow \emptyset$ procedure Finalize() 10 ret $\exists (K_i, N, A, M, C), (K_j, N', A', M', C') \in S$ s.t. 11 $(M \neq \perp \wedge M' \neq \perp) \wedge$ 12 $(C = C') \wedge$ 13 $(K_i, N, A, M) \neq (K_j, N', A', M') \wedge$ 14 $(\text{chk}(K_i, K_j) \vee \text{chk}(K_j, K_i))$ $\text{chk}(K_i, K_j)$ 16 if $t = 00 \wedge K_i \notin K_c \cup K_r \wedge K_j \notin K_c \cup K_r$ then ret 1 17 if $t = 01 \wedge K_i \notin K_r \cup K_c \wedge K_j \notin K_c$ then ret 1 18 if $t = 0X \wedge K_i \notin K_r \cup K_c$ then ret 1 19 if $t = 11 \wedge K_i \notin K_c \wedge K_j \notin K_c$ then ret 1 1A if $t = 1X \wedge K_i \notin K_c$ then ret 1 1B if $t = XX$ then ret 1 1C ret 0	procedure ENC(i, N, A, M) 20 if $K_i \notin K_r \cup K_c \wedge N \in N_i$ 21 then ret \perp 22 $C \leftarrow \Pi.E(K_i, N, A, M)$ 23 $N_i \leftarrow \{N\}$ 24 $S \leftarrow \{(K_i, N, A, M, C)\}$ 25 ret C procedure DEC(i, N, A, C) 30 $M \leftarrow \Pi.D(K_i, N, A, C)$ 31 $S \leftarrow \{(K_i, N, A, M, C)\}$ 32 ret M procedure REV(i) 40 $K_r \leftarrow \{K_i\}; \text{ret } K_i$ procedure COR(i, K) 50 $K_i \leftarrow K; K_c \leftarrow \{K_i\}$
---	--

Fig. 1: **The CAE-security game.** The encryption, decryption, reveal, and corruption oracles are on the right. On the left, the game finalization procedure depends on the collision check function chk , which in turn relies on the collision type parameter t of the game. This function places restrictions on the keys that the adversary may win with.

tional framework can be tuned to consider these collisions and more, allowing flexibility when using the definition to model real systems.

Our definition considers the strongest level of misattributions. That is, we require that $(K_i, N, A, M) \neq (K_j, N', A', M')$. This means the adversary wins as long as one of the inputs to encryption differ when creating the colliding ciphertext. We call a cAE scheme that attends to all encryption inputs *fully committing*.

Most other works only consider sub-tuples. For example, the notion of key commitment from Albertini et al. only requires that $K_i \neq K_j$ from the adversary when it creates a collision [2]. They show that key commitment is important for several real world systems. Nonetheless, this definition does not capture colliding ciphertexts that are generated under the same key. In section 4, we give a

transform as efficient as theirs while protecting against misattributions over the entire tuple of encryption inputs. (In a way, our transform is more efficient as it does not need to re-key every encryption call).

Contemporary work from Bellare and Hoang considers fully committing cAE schemes as well as sub-tuples [5]. Their argument for fully committing schemes is to provide ease of use. Prior definitions required different inputs to be committed to achieve the different security goals demanded by their relevant applications. The designer of an application may not know exactly what they need to be committed. So, if full commitment is inexpensive, then one should aim to do so.

Nonetheless, we provide an alternative CAE-security game that considers weaker misattributions, which we present in Appendix A. It uses an additional parameter allowing the specification of which encryption inputs are important when considering misattributions. However, we do note our construction CTX presented in Section 4 achieves full commitment efficiently.

RELATIONSHIP WITH NAE SECURITY. Previously, DGRW18 show how to construct a ciphertext for AES-GCM in such a way that it decrypts validly under two different keys [8]. This shows that nae-secure schemes do not achieve CAE_{xx}-security. In fact, the attack presented by DGRW18 does not require the adversary to have full control over the keys; it is possible to do the attack with only knowledge of the keys, which means nae-secure schemes do not achieve CAE₁₁-security either.

We show that naE schemes that are nae-secure in the multi-key sense—nae*-secure, are already CAE₀₀-secure. That is, when an adversary may not affect the keys in any way, it is already difficult to find colliding ciphertexts for nae-secure schemes.

Theorem 1. *Any authenticated encryption scheme Π that is nae*-secure is also CAE₀₀-secure. That is, for any adversary \mathcal{A} attacking the CAE₀₀-security of Π , there exists an adversary \mathcal{B} against the nae*-security of Π such that*

$$\mathbf{Adv}_\Pi^{\text{cae-00}}(\mathcal{A}) \leq 3 \cdot \mathbf{Adv}_\Pi^{\text{nae}^*}(\mathcal{B}) + \frac{q_e^2}{2^{\tau+1}}$$

where q_e is the number of encryption queries made by \mathcal{A} and τ is the expansion of the scheme Π . Furthermore, \mathcal{B} makes the same number of encryption and decryption queries that \mathcal{A} makes. That is, \mathcal{B} makes $\Theta(q_e)$ encryption queries and $\Theta(q_d)$ decryption queries where q_d is the number of decryption queries made by \mathcal{A} .

Proof. Let \mathcal{A} be an adversary attacking the CAE₀₀-security of Π . We assume that \mathcal{A} is nonce-respecting and that it does not query output of encryption to decryption as it would already know the answers of those queries and those queries would not help \mathcal{A} in obtaining a win. We also assume that \mathcal{A} never calls the reveal or corruption oracles as it can only win with a collision on a pair of honest keys. We can construct an adversary \mathcal{B} attacking the nae-security of Π as follows. Adversary \mathcal{B} sets up the CAE game as described in Fig. 1, maintaining its own set S to keep track of query and response tuples. Whenever \mathcal{A} makes

encryption or decryption queries, \mathcal{B} queries its own encryption and decryption oracles to provide a response for \mathcal{A} . When \mathcal{A} terminates, \mathcal{B} checks S to see if \mathcal{A} has created a colliding ciphertext. If it has, then \mathcal{B} returns 1. Otherwise, it returns 0.

Consider the three different ways that \mathcal{A} can add a winning ciphertext and its associated explanations to S . Either they were added through two decryption queries, an encryption and a decryption query, or two encryption queries. Let E_1, E_2, E_3 be the events that those yielded \mathcal{A} a win respectively. As these three events are all the ways to win, \mathcal{A} 's advantage is upper-bounded by the sum of the probabilities that each of these occur.

We bound the probabilities of each event by examining what happens when \mathcal{B} 's oracles are real or fake. For E_1 it is impossible for a winning tuple to be added to S when \mathcal{B} 's decryption oracle is fake as that oracle only ever returns \perp and a winning tuple must have a valid message. As such, \mathcal{B} only ever returns 0. However, if \mathcal{B} 's oracle is real, then \mathcal{B} will return 1. Hence, $\Pr[E_1] \leq \mathbf{Adv}_{\mathcal{B}}^{\text{nae}}$.

For E_2 a fake decryption oracle for \mathcal{B} makes winning through this event impossible by the same reasoning as that of E_1 , meaning \mathcal{B} only returns 0 here as well. Similarly, \mathcal{B} will return 1 if its oracles are real in this event. As such, the probability follows: $\Pr[E_2] \leq \mathbf{Adv}_{\mathcal{B}}^{\text{nae}}$.

For E_3 , \mathcal{B} can return 1 with a fake encryption oracle so long as \mathcal{A} gets a collision through its encryption queries. We can get the probability that this occurs by a birthday bound on the number of encryption queries made by \mathcal{A} . The birthday bound is over the random ciphertexts generated by the fake encryption oracle. With a real encryption oracle, \mathcal{B} always returns 1. This gives the probability $\Pr[E_3] \leq \mathbf{Adv}_{\mathcal{B}}^{\text{nae}} + \frac{q_e^2}{2^{r+1}}$. □

4 The CTX Construction

THE CTX SCHEME. Recall that a cAE scheme is *fully committing* if it commits to the key, nonce, AD, and message and not some subset of them. We say that a scheme is *efficient* if its cost of getting cAE security on top of nAE security is independent of the message length. We call a scheme *strong* if it achieves CAE_{xx} security. Our CTX construction is fully committing, efficient, and strong.

Let $\Pi = (\mathcal{E}, \mathcal{D})$ be a tag-based nAE scheme. That is, ciphertexts it outputs consist of a ciphertext core C and an authentication tag T . We assume that the encryption algorithm \mathcal{E} can be split into two independent algorithms \mathcal{E}_1 and \mathcal{E}_2 such that on inputs K, N, A, M , \mathcal{E}_1 produces the core C and \mathcal{E}_2 outputs the tag T . The core C is the same length as M . As such, \mathcal{E}_1 is bijective when K, N, A are fixed. The inverse of \mathcal{E}_1 is then decryption's subroutine \mathcal{D}_1 , which takes in K, N, A and just the core C , and outputs M . That is, $\mathcal{D}_1(K, N, A, \mathcal{E}_1(K, N, A, M)) = M$. Common schemes like GCM and OCB satisfy these structural demands.

From such an nAE scheme Π and a collision-resistant hash function H , we can construct a CAE_{xx}-secure cAE scheme, $\text{CTX}[\Pi, H]$. CTX's main mechanism

$\text{CTX}.\mathcal{E}(K, N, A, M)$	$\text{CTX}.\mathcal{D}(K, N, A, C)$
20 $C \leftarrow \Pi.\mathcal{E}_1(K, N, A, M)$	30 $C \parallel T \leftarrow \mathcal{C}$
21 $T \leftarrow \Pi.\mathcal{E}_2(K, N, A, M)$	31 $M \leftarrow \Pi.\mathcal{D}_1(K, N, A, C)$
22 $T^* \leftarrow H(K, N, A, T)$	32 $T' \leftarrow \Pi.\mathcal{E}_2(K, N, A, M)$
23 ret $C \parallel T^*$	33 if $T \neq H(K, N, A, T')$ then ret \perp
	34 ret M

Fig. 2: A CAE_{XX}-secure cAE scheme $\text{CTX}[\Pi, H]$ built from a tag-based nAE scheme Π and a collision-resistant hash function H . The nAE encryption and decryption algorithms can be broken down into \mathcal{E}_1 , \mathcal{E}_2 , and \mathcal{D}_1 . These create the ciphertext core, create the authentication tag, and recover the message from the core respectively.

is hashing the authentication tag T along with K, N, A into a new tag T^* . This effectively makes T^* function as the nAE authenticity check and a commitment to K, N, A . The name CTX captures the scheme’s ciphertext structure, which is a ciphertext core followed by a modified tag. The ‘X’ in the name suggests the scheme’s XX-security level. The scheme is presented in Fig. 2.

We claim that CTX is CAE_{XX}-secure as long as H is collision-resistant.

Theorem 2. *Let $\Pi = (\mathcal{E}, \mathcal{D})$ be a tag-based nAE scheme and let H be a collision-resistant hash function. Let $\mathcal{E}_1, \mathcal{E}_2, \mathcal{D}_1$ be the algorithms used by Π to encrypt messages into ciphertext cores, create authentication tags, and decrypt cores into messages respectively. Let $\text{CTX}[\Pi, H]$ be an nAE scheme constructed from Π and H as described in Fig. 2. Then, for any adversary \mathcal{A} attacking the CAE_{XX}-security of Π' , there exists an adversary \mathcal{B} , explicitly given in the proof of this theorem and depending only on \mathcal{A} as a black-box, such that*

$$\mathbf{Adv}_{\text{CTX}}^{\text{cae-XX}}(\mathcal{A}) \leq \mathbf{Adv}_H^{\text{col}}(\mathcal{B}).$$

Proof. We construct adversary \mathcal{B} to find a winning collision for H as follows. Adversary \mathcal{B} sets up the CAE game and runs \mathcal{A} , answering its queries appropriately. When \mathcal{A} terminates, it will have produced a pair of winning tuples for the CAE game: $(K_i, N, A, M, C \parallel T^*)$, $(K_j, N', A', M', C \parallel T^*)$. Then, \mathcal{B} can compute $T = \Pi.\mathcal{E}_2(K_i, N, A, M)$ and $T' = \Pi.\mathcal{E}_2(K_j, N', A', M')$ to produce authentication tags for the winning tuples. Furthermore, it must be the case that $T^* = H(K_i, N, A, T) = H(K_j, N', A', T')$ as that is how $C \parallel T^*$ of the winning tuples was produced in the first place.

So for $(K_i, N, A, T), (K_j, N', A', T')$ to be a winning collision for \mathcal{B} , the two tuples must not be equivalent. Suppose for contradiction that they are equivalent. Then for \mathcal{A} ’s tuples to have won the CAE game it must be the case that $M \neq M'$. But this is impossible as it would violate the bijectivity of \mathcal{E}_1 . Since core C is fixed and $(K_i, N, A) = (K_j, N', A')$ are fixed from the collision, there exists only one M'' such that $\mathcal{E}_1(K_i, N, A, M'') = C = \mathcal{E}_1(K_j, N', A', M'')$. Thus $M \neq M'$

is a contradiction and $(K_i, N, A, T) \neq (K_j, N', A', T')$ follows. In conclusion, the winning collision for \mathcal{B} is $H(K_i, N, A, T) = H(K_j, N', A', T') = T^*$. \square

From Theorem 2, we see that the CAE_{XX}-security of CTX is bounded by the collision-resistance of the hash function it employs. One can break this with about $2^{\mu/2}$ operations doing a birthday-attack, which is why we recommend having CTX tag length be 160-bits over, say 128-bits. This raises a question: Can one lower the security requirement (something weaker than CAE_{XX}) and avoid the birthday bound?

The answer is that with CTX, you cannot unless there are further assumptions made of the nAE scheme Π that it uses. There exists an attack on the CAE_{OX}-security of CTX using a birthday attack under standard assumptions on Π . We explicitly describe this attack in Appendix C.

It remains to show that CTX remains nAE-secure after its transform. We do so in the random oracle model (ROM), denoting CTX as CTX[Π] (as opposed to CTX[Π, H]) when it is in the ROM. The privacy and authenticity notions in Theorem 3 can be attained by breaking the nAE security notion in 2 apart and can be found in [19]. We prove Theorem 3 in Appendix B.

Theorem 3. *Let $\Pi = (\mathcal{E}, \mathcal{D})$ be a tag-based nAE scheme with an expansion of τ . Let CTX[Π] be the scheme described in Fig. 2. Fix an integer $\delta \geq 0$. Then, in the random oracle model, for any adversary \mathcal{A}_1 attacking the privacy of CTX, we can construct nonce-respecting (explicitly given) adversaries \mathcal{B}_1 and \mathcal{B}_2 attacking the privacy of Π such that*

$$\mathbf{Adv}_{\text{CTX}[\Pi]}^{\text{priv}}(\mathcal{A}_1^H) \leq \mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{B}_1) + \mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{B}_2) + \frac{q_H}{2^{\delta+\tau}}$$

where q_H is the number of random oracle queries made by \mathcal{A}_1 . Let q_e be the number of encryption queries made by \mathcal{A}_1 . Then \mathcal{B}_1 also makes q_e queries to its own encryption oracle and \mathcal{B}_2 makes $q_e + 1$ such queries.

Furthermore, for any adversary \mathcal{A}_2 attacking the authenticity of CTX, there exists adversary \mathcal{B}_3 attacking the authenticity of Π with advantage

$$\mathbf{Adv}_{\text{CTX}[\Pi]}^{\text{auth}}(\mathcal{A}_2^H) \leq \mathbf{Adv}_{\Pi}^{\text{auth}}(\mathcal{B}_3) + \frac{1}{2^\mu}$$

where μ is the output length of the random oracle. We give \mathcal{B}_3 explicitly in the proof. If \mathcal{A}_2 makes q_e encryption oracle queries, then \mathcal{B}_3 makes $q_e + 1$ queries to its own oracle.

Note that the last term in the privacy bound $\frac{q_H^2}{2^{\delta+\tau}}$ can be made small by choice of δ , so it will not result in much loss.

5 Commitment Security of GCM and OCB

Prior work from GLR17 and DGRW18 has shown that it is possible to construct a colliding ciphertext with GCM when the attacker has control of both keys

$\text{NAE.E}(K, N, A, M)$	$\text{NAE.D}(K, N, A, C)$
00 $K_1 \parallel K_2 \leftarrow K; P \leftarrow G(K_1, N)$	10 $K_1 \parallel K_2 \leftarrow K; C \parallel T \leftarrow C$
01 $C \leftarrow M \oplus P[0.. M -1]$	11 $\text{if } H(K_2, C \parallel A) \neq T \text{ then ret } \perp$
02 $T \leftarrow H(K_2, C \parallel A)$	12 $P \leftarrow G(K_1, N)$
03 $\text{ret } C \parallel T$	13 $M \leftarrow C \oplus P[0.. M -1]; \text{ret } M$

Fig. 3: A simple nAE scheme given a PRG G and a MAC H . GCM has a comparable structure if one considers the counter-mode operations as the PRG and GHASH as the MAC.

[8,12]. DGRW18 mentions that with their attack, control over the keys is not necessary, only knowledge of the keys is. Here, we show that it is possible for an attacker to create a colliding ciphertext with knowledge of only one key. That is, there exists an attack that violates the CAE₀₁-security of GCM. As GCM is nae-secure, this attack means that nae-security does not imply CAE₀₁-security.

A SIMPLE NAE SCHEME. Before we present the attack, consider a simple nAE scheme $\text{NAE}[G, H]$ built on a PRG G and a MAC H . The definition of $\text{NAE}[G, H]$ is given in Fig. 3. In our pseudocode, we write $S[0..n]$ to denote a substring of the bitstring S starting from the 0th bit to the n th bit.

However, NAE is vulnerable to a variety of CAE attacks given that the MAC H is *targetable*. Suppose that the key K used for computing H is known and there exists an arbitrary target tag T that an adversary is interested in producing. We call H targetable if there exists a target function target that takes in K and T and outputs a message M such that $H(K, M) = T$. We say that H is *prefix-targetable* if there exists a prefixed target function may also take in an additional argument C , a prefix, such that $H(K, C \parallel M) = T$. GHASH, the MAC used by GCM, is prefix-targetable, and we will show how shortly.

ATTACK ON GCM. The simple nAE scheme described has structure similar to GCM. For concreteness, we assume the blockcipher GCM employs E has a block size of 128 bits. GCM uses E in counter mode with the nonce as part of the initial counter in order to generate a one-time pad. This acts like the PRG that the simple scheme uses. For a key K , GCM uses $K' = E_K(0^{128})$ when computing its MAC, GHASH. For a ciphertext C , the tag $T = H(K', A, C) \oplus E_K(N \parallel 0^{31})$ where A is the AD, C is the ciphertext, N is the nonce and H is the GHASH function. We follow the GCM specification of [15,16].

GHASH works by computing a polynomial over the field $\text{GF}(2^{128})$ using $E_K(0^{128})$ as the variable and the ciphertext and AD blocks as coefficients. By block, we mean blocks of b bits that can be used as input into a blockcipher. If the last block isn't a full 128 bits, GCM pads it with zeroes until it is. Let there be c ciphertext blocks and a AD blocks in the ciphertext and AD. Let len be a function where given some input, it outputs a 64-bit representation of said

input. Let $P = E_{K'}(0^{128})$. Then GHASH is computed as follows (addition and multiplication done over $\text{GF}(2^{128})$):

$$\text{GHASH}(K', A, C) = \left[\sum_{i=1}^a A_i \cdot P^{a+c+2-i} \right] + \left[\sum_{i=1}^c C_i \cdot P^{c+2-i} \right] + (\text{len}(A) \parallel \text{len}(C)) \cdot P \quad (1)$$

And the tag T is finalized as:

$$T = \text{GHASH}(K', A, C) \oplus E_{K'}(N \parallel 0^{31}1) \quad (2)$$

where N is the nonce.

Observe that the entire MAC is prefix-targetable as if one knows K' and T , one can compute $A = \text{ptarget}(K', T, C)$ for a ciphertext C by evaluating the polynomial. Explicitly, we can solve for a single block AD A as follows:

$$A = \left[T \oplus E_{K'}(N \parallel 0^{31}1) + (\text{len}(A) \parallel \text{len}(C)) \cdot P + \left[\sum_{i=1}^c C_i \cdot P^{c+2-i} \right] \right] \cdot (P^{c+2})^{-1} \quad (3)$$

Once we can compute A with the prefix-targeting function, we have an CAE_{01} attack, call it adversary \mathcal{A} , as follows: \mathcal{A} selects an arbitrary nonce N and AD A ; \mathcal{A} queries its encryption oracle, asking for the encryption of a string of 0's of length m under K_0 , N , and A , and receives $C \parallel T$ back; \mathcal{A} uses its reveal oracle to learn K_1 ; \mathcal{A} computes a one-time pad P in the style of GCM using K_1 and the nonce N ; \mathcal{A} computes a message M' as the xor of P and C ; \mathcal{A} uses the prefix-target function $\text{ptarget}(K', T, C)$ as shown in Equation 3 where K' is the blockcipher E applied to 0^{128} with K_1 (how GHASH is keyed) and acquires an AD A' ; \mathcal{A} queries its encryption oracle with K_1, N, A', M' and receives a winning collision on $C \parallel T$. This attack on GCM prove that nae-security does not imply even CAE_{01} -security as GCM is nae-secure.

While Equation 3 computes a single AD block that allows us to obtain a target tag, this is actually not restrictive. An attack can use an arbitrary AD A , perhaps with actually relevant header information, and search for a single block A' that they can add to that AD to satisfy the equation. To capture this in terms of prefix-targeting, one would compute $A' = \text{ptarget}(K', T, A \parallel C)$ instead of $\text{ptarget}(K', T, C)$. Keep in mind that the dummy block can be placed anywhere in A , but we limit our description to prepending for simplicity.

ATTACK ON OCB. We now turn to performing a CAE_{01} attack on OCB. We follow the specification of OCB as described in [13].

During encryption, OCB computes an offset Δ for each message block using the key and the nonce. This offset is xor-ed with each message block before being processed by a blockcipher under the key. The output of the blockcipher is then

```

 $\mathcal{A}_{\text{OCB},01}^{\text{CAE}}$ 
20  $N \leftarrow \mathcal{N}; M \leftarrow \{0,1\}^m; C \parallel T \leftarrow \text{ENC}(0, N, \varepsilon, M)$ 
21  $C_1 \parallel C_2 \parallel \dots \parallel C_n \leftarrow C$  where  $|C_i| = b$  for all  $i \in [1..n]$ 
22  $K \leftarrow \text{REV}(1); \Delta \leftarrow \text{init}(K, N)$ 
23 for  $i = 1$  to  $n$  do
24    $\Delta \leftarrow \text{incr}_i(\Delta)$ 
25    $P \leftarrow C_i \oplus \Delta; M'_i \leftarrow E_K^{-1}(P) \oplus \Delta$ 
26    $\text{chk} \leftarrow M'_1 \oplus M'_2 \oplus \dots \oplus M'_n$ 
27    $F \leftarrow \text{chk} \oplus \text{incr}_\$ (\Delta); F \leftarrow E_K(F)$ 
28    $\Delta \leftarrow \text{incr}_1(0^{128})$ 
29    $auth \leftarrow T \oplus F; A \leftarrow E_K^{-1}(\text{auth}) \oplus \Delta$ 
2A    $\text{ENC}(1, N, A, M'_1 \parallel \dots \parallel M'_n)$ 

```

Fig. 4: An CAE₀₁ attack on OCB. For simplicity, this attack assumes that the length of the ciphertext is a multiple of the blockcipher E 's block size b . The attack written here shows how to compute a single associated data block to get a colliding ciphertext (line 26-29). But it should be noted that it is possible to mount the attack by choosing an arbitrary AD first and computing a single block that satisfies a necessary value for $auth$ to get the colliding tag.

xor-ed with the offset again, finalizing a ciphertext block. Since the adversary has a revealed key K_j and a nonce of its choice N , it can freely compute the offsets for each block. This allows it to decrypt the target colliding ciphertext $C \parallel T$, where C is the ciphertext core and T is the authentication tag, into some message M . The next step requires the adversary to ensure that T verifies for M under K_j and N . In OCB, tag T is generated first with a checksum that consists of an xor over all message blocks. The adversary can do this over the bogus message M it got from decryption. This checksum is then xor-ed with a special offset, again computable with knowledge of key and nonce, before being processed by the blockcipher. This output, F , is xor-ed with a block called “ $auth$ ” which finalizes the tag T .

The adversary then needs $auth = F \oplus T$ for T to remain valid with K_j, N, M, C . To do so, it has a choice of AD A . OCB computes $auth$ by computing offsets for each block of AD, xor-ing the offsets and blocks together, and applying the blockcipher on the results (a process identical to how the message blocks are processed with the exception of how the offsets are initialized). Each of these blocks are then xor-ed with each other, finalizing a single block $auth$. To acquire an A that finishes the attack, the adversary deciphers $F \oplus T$ with the blockcipher, xors the result with the appropriate offset, and uses that for its final query $\text{ENC}(j, N, A, M)$. The attack is described in code in Fig. 4.

Like the attack on GCM, it should be noted that the attack is not limited to a single AD block. An adversary may select an arbitrary AD A' that it wants

to use to mount the attack. It can then compute a single dummy block B' to append to the end of A' to create an AD that validates decryption. Specifically, the adversary first computes F as described above. Then it computes a value $auth'$ over the blocks of A' . The value $B = auth' \oplus F \oplus T$ will then be the “enciphered” AD block corresponding to B' . So the adversary only needs to decipher B and apply the appropriate offset to compute B' . The final AD it uses is $A = A' \parallel B'$.

6 Other Committing AE Notions

Here we describe the committing AE notions of previous authors and highlight their differences. A summary of each of their definitions can be found in Table 1. The first to study committing encryption in the AE setting was Farshim, Orlandi, and Rošie (FOR17) [10]. Calling the property *key-robustness*, FOR17 give a set of definitions capturing different adversarial behaviors that can result in the misattribution of a ciphertext. Their strongest notion, *full robustness*, requires an adversary to produce two keys and a ciphertext (K, K', C) such that C decrypts validly under both keys. It needs to be noted that FOR17 study randomized AE without AD support.

An interesting result from FOR17 is that security for such AE schemes implies *semi-full robustness*. In this notion, two keys are generated uniformly at random and one of them is shown to the adversary. With the help of encryption and decryption oracles for the hidden key, the adversary must find a ciphertext that validly decrypts under both keys. This definition is comparable to our CAE₀₁ notion, where the adversary must find a misattribution with a revealed and an honest key. One of our results is the existence of CAE₀₁-attacks against AES-GCM and OCB, which implies that nAE security does not grant 01-security. This seemingly contradicts FOR17’s result of semi-full robustness because their analysis is for AE schemes without AD support.

In the same year as FOR17, Grubbs, Lu, and Ristenpart (GLR17) study message franking, which as they describe it, is the *verifiable* reporting of abusive messages in encrypted messaging systems [12]. To accomplish this goal, they use committing AE, focusing on randomized AE with AD support (AEAD) as it is more applicable to current encrypted messaging systems. In their model, there is a sender, a receiver, and a third party that verifies abuse reports. Every ciphertext comes with a commitment tag that serves as a commitment to the message and AD. Decryption produces an opening for the commitment alongside recovering the message. Their committing AE notion adds an additional verification algorithm as it is the third party’s role to verify the commitment using that opening. We conflate their decryption and verification algorithms for ease of discussion and comparison to other notions.

There are several parts of their committing AE notion that make it difficult to compare as they tend to other things besides preventing misattributions. The part that attends to misattributions is their notion of *receiver binding*. This notion asks the adversary to find a ciphertext C and two tuples

Paper	AE Variant	Committing AE Definition
FOR17 [10]	Probabilistic AE, No AD support	<i>Full robustness</i> - \mathcal{A} finds (K, K', C) s.t. decryption of C with both keys is successful.
GLR17 [12]	Probabilistic and deterministic AEAD	<i>Receiver-binding</i> - \mathcal{A} finds $((K, A, M), (K', A', M'), C)$ s.t. decryption of C with both sub-tuples is successful and $(A, M) \neq (A', M')$.
DGRW18 [8]	Probabilistic AEAD	<i>Strong Receiver-binding</i> - Same as receiver-binding except $(K, A, M) \neq (K', A', M')$.
ADGKLS20 [2]	Deterministic AEAD	<i>Key-commitment</i> - \mathcal{A} finds $((K, N, A, M), (K', N, A', M'), C)$ through ENC, DEC queries s.t. $K \neq K'$ and $M, M' \neq \perp$.
BH22 [5]	Deterministic AEAD + misuse-resistant AE	<i>CMT(D)_s-ℓ security</i> - \mathcal{A} finds $(K_1, N_1, A_1, M_1), \dots, (K_s, N_s, A_s, M_s)$ such that what is committed from each tuple is distinct. The parameter ℓ specifies “what is committed.”
This Paper	Deterministic AEAD	<i>CAE_t-security</i> - \mathcal{A} finds $((K, N, A, M), (K', N', A', M'), C)$ through ENC, DEC queries s.t. $(K, N, A, M) \neq (K', N', A', M')$ and $M, M' \neq \perp$. The parameter t specifies how \mathcal{A} interacts with the keys.

Table 1: A comparison of the subtly different definitions in CAE literature.

$(K, A, M), (K', A', M')$ such that decrypting C with those keys and ADs results in those (valid) messages. The adversary must do so in a way such that $(A, M) \neq (A', M')$. This definition does not prevent the possibility of an adversary finding two keys that can validly decrypt C into M using A .

Dodis, Grubbs, Ristenpart, and Woodage (DGRW18) [8] extend GLR17’s receiver binding to *strong receiver binding*. This notion accounts for the key to address the way receiver binding does not. One can argue that strong receiver binding commits to all encryption inputs for randomized AEAD. As a building block for cAE, DGRW18 introduce a new primitive *encryptment* that serves as a one-time use, deterministic encryption and commitment of a message.

One goal that GLR17 and DGRW18 consider that other works do not (including ours) is that of *multiple-opening security*. This security notion allows

different ciphertexts encrypted under the same key to be “opened” and verified without jeopardizing the security of unopened ciphertexts. This is particularly useful in the message franking context as it allows a receiver to report a ciphertext to the verifying party without having to reveal the secret key, which would ruin the security of all other ciphertexts sent under that key.

Working with deterministic AEAD, Albertini, Duong, Gueron, Kölbl, Luykx, and Schmieg (ADGKLS20) [2] define their security goal as *key-commitment*. The adversary, in this notion, is tasked with finding a ciphertext C and two “explanations” $(K, N, A, M), (K', N, A', M')$ such that the messages are valid and $K \neq K'$.

Bellare and Hoang (BH22), in a contemporary work, target fully committing schemes [5]. They attend to deterministic AEAD as well as misuse-resistant AE with encryption inputs K, N, A, M . Their committing security notion is $\text{CMT}(\text{D})_{s,\ell}$ where s is an integer and $\ell \in \{1, 3, 4\}$. The presence or absence “D” denotes whether the adversary is tasked with finding multiple decryption inputs that validly decrypt the same ciphertext or multiple encryption inputs that encrypt to the same ciphertext. The parameter ℓ determines what is committed: $\ell = 1$ denotes just the key, $\ell = 3$ denotes everything but the plaintext, and $\ell = 4$ denotes full commitment. Comparatively, our cAE definition presented in Section 3 does not allow for tweaking for commitments of sub-tuples of inputs, but the alternative framework given in Appendix A does. The s parameter generalizes their definition to capture misattributions with more than two valid explanations—what they call *multi-input committing security*. That is, s is the number of distinct (K, N, A, M) tuples the adversary needs to find that encrypt to the same C . While $s = 2$ implies all $s \geq 2$, Bellare and Hoang motivate this dimension of their definition by giving schemes where bounds on adversarial advantage improve as s grows. All in all, they are the first to study misuse-resistant AE and multi-input committing security in the cAE space.

CONSTRUCTIONS. We describe a number of selected constructions from the above works. These constructions, each satisfying the committing AE notion defined in the work of their origin, are presented in Table 2.

Recall that FOR17 are in the probabilistic AE setting without associated data. Their construction $\text{EtM}[\mathcal{E}, H]$ creates a tag that provides authenticity while serving as a commitment to the encryption key as well. This is comparable to how CTX’s tag provides authenticity while committing to all of K, N, A, T .

The scheme $\text{CEP}[G, F, F^{\text{cr}}]$ is the deterministic AEAD construction from GLR17. It makes two passes over the message—one to encrypt it one-time-pad-style using output from the PRG G and the other to commit to the message and AD using the collision-resistant PRF F^{cr} . The ciphertext output is expanded by both a tag for authenticity and a commitment—the output lengths of the two PRFs. Comparatively, our CTX construction requires no passes over the message and would typically expand ciphertexts from a 128-bit authentication tag to a 160-bit hash function output that gives both cAE security and nAE authenticity. An advantage of CEP is that one can verify the commitment without revealing

Construction	Description
$\text{EtM}[\mathcal{E}, H]$ [10]	\mathcal{E} is AE scheme, H is CR MAC. Encrypt M w/ \mathcal{E} under key K_e to get C . Get T by using MAC w/ key K_h on (C, K_e) . Output $C \parallel T$.
$\text{CEP}[G, F, F^{\text{cr}}]$ [12]	G is PRG. F, F^{cr} are PRFs. F^{cr} is CR Use G w/ K and N to get K_0, K_1, P . Use $P \oplus M$ to get C_1 . Use F^{cr} w/ K_0 on A, M for C_2 . Use F w/ K_1 on C_2 to get T . Output $C_1 \parallel T \parallel C_2$.
HFC^* [8]	HFC is an <i>encryptment</i> scheme built from a compression function and a padding scheme. DGRW18 show a simple transform that promotes an encryption scheme into a cAE scheme.
CommitKey_{IV} [2]	\mathcal{E} is nAE scheme. F_0, F_1 are independent CR PRFs. Get K_e from using F_0 w/ K on nonce N . Get K_c from using F_1 w/ K on nonce N' . Use \mathcal{E} on N, A, M to get C . Output $C \parallel K_c$.
$\text{UtC}[\mathcal{E}, F]$ [5]	\mathcal{E} is nAE scheme. F is <i>committing</i> PRF. Get (P, L) from $F(K, N)$. Get C from $\mathcal{E}(L, N, A, M)$. Output $P \parallel C$.
$\text{HtE}[\mathcal{E}, H]$ [5]	\mathcal{E} is a <u>CMT-1</u> nAE scheme. H is a CR function. Get L from $H(K, (N, A))$. Output $\mathcal{E}(L, N, \varepsilon, M)$.

Table 2: A comparison of selected constructions targeting their respective cAE security goals. *Not a committing AE scheme, but closely related.

the encryption key. One only needs to reveal K_0 to do so. This is in line with GLR17’s additional goal of multiple opening security.

DGRW18 had similar goals to GLR17 as they both investigated committing AE for the purpose of message franking. They propose a new primitive, encryptment, that we do not describe in detail here. Encryptment is a primitive that simultaneously encrypts and commits a message and is one-time use. They give a concrete encryptment scheme HFC that uses a compression function and a padding scheme. They give a simple transform that builds a cAE scheme out of an encryption scheme and a probabilistic nAE scheme. We note that HFC requires a pass over the message to apply encrypt and commit it.

The CommitKey scheme from ADGKLS20 comes in four flavors. We describe the variant CommitKey_{IV} here. It consists of an nAE scheme and two independent collision-resistant PRFs. On encryption, the PRFs are used on the nonce to generate an encryption key and a “key-commitment.” The encryption key is then used to perform routine nAE encryption on the message, producing a ciphertext. Encryption returns both the ciphertext and the key-commitment. It commits only to the key and as such, does not require any passes over the plaintext.

With this scheme, it is possible to find a misattribution where different AD lead to valid decryptions.

One can argue that this kind of misattribution may not be impactful to real-world systems. But **CTX** protects against these misattributions as well and without giving up efficiency. In fact, **CTX** enjoys the efficiency benefit of not having the re-key with each message encrypted.

That argument is specious, in any case. It is difficult for designers of systems to know exactly what needs to be committed to achieve their security goals. GLR17 and DGRW18 showed message franking requires the commitment of the header and message. ADGKLS20 found that various real-world systems (key management services, envelope encryption, and Subscribe with Google [3]) had potential vulnerabilities from lack of key commitment. It is not always clear what exactly needs to be committed, so a scheme that can inexpensively commit to everything would provide a way to cover all bases for application designers.

Bellare and Hoang give a fully committing cAE construction that builds off of one that only commits to a key. Their UtC construction only commits to the key (CMT-1 secure going by their notions). It uses a primitive they call a *committing PRF* which informally outputs a commitment to the key and the PRF input along with the conventional PRF output. They describe an efficient committing PRF in their paper.

To promote a CMT-1 secure scheme to a fully committing (CMT-4) one, BH22 give the HtE transform. Like our **CTX** construction, the application of HtE to UtC commits to everything without having to make a pass over the plaintext beyond encrypting it. The ciphertext expansion of BH22’s transform however is expected to be at least 128-bits—the block length of the blockcipher that their committing PRF employs. On the other hand, **CTX** is expected to replace a conventional nAE tag, say 128-bits, to a 160-bit tag that provides both nAE authenticity and the commitment to all encryption inputs. This would be a 32-bit expansion compared to the expansion by a full block.

Acknowledgements

The authors gratefully acknowledge insightful input from Mihir Bellare and Tom Shrimpton. Thanks to Mihir specifically for the discussion on **CTX**’s security level with notions weaker than CAE_{xx}-security as well as **CTX**’s privacy and authenticity.

This work was funded by NSF grant CNS 1717542. Many thanks to the NSF for their years of support.

References

1. M. Abdalla, M. Bellare, and G. Neven. Robust encryption. In D. Micciancio, editor, *TCC 2010: 7th Theory of Cryptography Conference*, volume 5978 of *Lecture Notes in Computer Science*, pages 480–497, Zurich, Switzerland, Feb. 9–11, 2010. Springer, Heidelberg, Germany. 3

2. A. Albertini, T. Duong, S. Gueron, S. Kölbl, A. Luykx, and S. Schmieg. How to abuse and fix authenticated encryption without key commitment. Cryptology ePrint Archive, Report 2020/1456, 2020. <https://eprint.iacr.org/2020/1456>. 1, 2, 3, 6, 7, 16, 17, 18, 21
3. J. Albrecht. Introducing subscribe with google, Mar 2018. <https://blog.google/outreach-initiatives/google-news-initiative/introducing-subscribe-google/>. 19
4. M. Bellare. A concrete-security analysis of the apple PSI protocol, July 2021. https://www.apple.com/child-safety/pdf/Alternative_Security_Proof_of_Apple_PSI_System_Mihir_Bellare.pdf. 6
5. M. Bellare and V. T. Hoang. Efficient schemes for committing authenticated encryption. In O. Dunkelman and S. Dziembowski, editors, *Advances in Cryptology – EUROCRYPT 2022, Part II*, volume 13276 of *Lecture Notes in Computer Science*, pages 845–875, Trondheim, Norway, May 30 – June 3, 2022. Springer, Heidelberg, Germany. 4, 8, 16, 17, 18, 21
6. M. Bellare and P. Rogaway. The security of triple encryption and a framework for code-based game-playing proofs. In S. Vaudenay, editor, *Advances in Cryptology – EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*, pages 409–426, St. Petersburg, Russia, May 28 – June 1, 2006. Springer, Heidelberg, Germany. 24
7. M. Bellare and B. Tackmann. The multi-user security of authenticated encryption: AES-GCM in TLS 1.3. In M. Robshaw and J. Katz, editors, *Advances in Cryptology – CRYPTO 2016, Part I*, volume 9814 of *Lecture Notes in Computer Science*, pages 247–276, Santa Barbara, CA, USA, Aug. 14–18, 2016. Springer, Heidelberg, Germany. 5
8. Y. Dodis, P. Grubbs, T. Ristenpart, and J. Woodage. Fast message franking: From invisible salamanders to encryption. In H. Shacham and A. Boldyreva, editors, *Advances in Cryptology – CRYPTO 2018, Part I*, volume 10991 of *Lecture Notes in Computer Science*, pages 155–186, Santa Barbara, CA, USA, Aug. 19–23, 2018. Springer, Heidelberg, Germany. 1, 3, 6, 8, 12, 16, 18, 21
9. P. Farshim, B. Libert, K. G. Paterson, and E. A. Quaglia. Robust encryption, revisited. In K. Kurosawa and G. Hanaoka, editors, *PKC 2013: 16th International Conference on Theory and Practice of Public Key Cryptography*, volume 7778 of *Lecture Notes in Computer Science*, pages 352–368, Nara, Japan, Feb. 26 – Mar. 1, 2013. Springer, Heidelberg, Germany. 3
10. P. Farshim, C. Orlandi, and R. Rosie. Security of symmetric primitives under incorrect usage of keys. Cryptology ePrint Archive, Report 2017/288, 2017. <https://eprint.iacr.org/2017/288>. 1, 2, 3, 6, 15, 16, 18, 21
11. Y. Gertner and A. Herzberg. Committing encryption and publicly-verifiable sign-encryption. Cryptology ePrint Archive, Report 2003/254, 2003. <https://eprint.iacr.org/2003/254>. 3
12. P. Grubbs, J. Lu, and T. Ristenpart. Message franking via committing authenticated encryption. In J. Katz and H. Shacham, editors, *Advances in Cryptology – CRYPTO 2017, Part III*, volume 10403 of *Lecture Notes in Computer Science*, pages 66–97, Santa Barbara, CA, USA, Aug. 20–24, 2017. Springer, Heidelberg, Germany. 1, 3, 6, 12, 15, 16, 18, 21
13. T. Krovetz and P. Rogaway. The software performance of authenticated-encryption modes. In A. Joux, editor, *Fast Software Encryption – FSE 2011*, volume 6733 of *Lecture Notes in Computer Science*, pages 306–327, Lyngby, Denmark, Feb. 13–16, 2011. Springer, Heidelberg, Germany. 13

14. J. Len, P. Grubbs, and T. Ristenpart. Partitioning oracle attacks. Cryptology ePrint Archive, Report 2020/1491, 2020. <https://eprint.iacr.org/2020/1491>. 1, 4
15. D. McGrew and J. Viega. The galois/counter mode of operation (gcm). *submission to NIST Modes of Operation Process*, 20:0278–0070, 2004. 12
16. D. A. McGrew and J. Viega. The security and performance of the Galois/counter mode (GCM) of operation. In A. Canteaut and K. Viswanathan, editors, *Progress in Cryptology - INDOCRYPT 2004: 5th International Conference in Cryptology in India*, volume 3348 of *Lecture Notes in Computer Science*, pages 343–355, Chennai, India, Dec. 20–22, 2004. Springer, Heidelberg, Germany. 12
17. C. Namprempre, P. Rogaway, and T. Shrimpton. Reconsidering generic composition. In P. Q. Nguyen and E. Oswald, editors, *Advances in Cryptology – EUROCRYPT 2014*, volume 8441 of *Lecture Notes in Computer Science*, pages 257–274, Copenhagen, Denmark, May 11–15, 2014. Springer, Heidelberg, Germany. 4
18. P. Rogaway. Formalizing human ignorance. In P. Q. Nguyen, editor, *Progress in Cryptology - VIETCRYPT 06: 1st International Conference on Cryptology in Vietnam*, volume 4341 of *Lecture Notes in Computer Science*, pages 211–228, Hanoi, Vietnam, Sept. 25–28, 2006. Springer, Heidelberg, Germany. 4
19. P. Rogaway, M. Bellare, J. Black, and T. Krovetz. OCB: A block-cipher mode of operation for efficient authenticated encryption. In M. K. Reiter and P. Samarati, editors, *ACM CCS 2001: 8th Conference on Computer and Communications Security*, pages 196–205, Philadelphia, PA, USA, Nov. 5–8, 2001. ACM Press. 11, 22

A CAE with Misattribution Types

MODELING WEAKER MISATTRIBUTIONS. The CAE definition presented in 3 only attends to the strongest level of misattributions. Specifically, a ciphertext C experiences the strongest level of misattribution if an adversary finds two distinct explanations $(K, N, A, M) \neq (K', N', A', M')$ for it. We call schemes that are resistant against this type of misattribution *fully committing*.

However, one may be interested in schemes that protect against weaker misattributions. After all, other works have investigated these weaker notions [10,12,8,2] and contemporary work by [5] give a framework that supports committing to subsets of the encryption inputs.

In addressing this, we have an alternative definition given in Fig. 5. In this definition, the game is parametrized with an additional four-bit string u . This string allows the specification of which elements of the (K, N, A, M) tuple one should commit to. The first bit corresponds to the key, the second the nonce, and so on. This framework is finer-grained than the contemporary framework of [5] as theirs only captures three kinds of commitments: committing to just the key; committing to the key, nonce, and AD; and committing to everything. The framework here allows the expression of all sixteen ways one can commit to encryption inputs.

<u>$\text{CAE}_{\Pi, \mathbf{t}, \mathbf{u}}$</u>	
procedure Finalize()	procedure Initialize()
10 ret $\exists (K_i, N, A, M, C), (K_j, N', A', M', C') \in \mathbb{S}$ s.t.	00 for $i \in \mathbb{N}$ do
11 $(M \neq \perp \wedge M' \neq \perp) \wedge$	01 $K_i \leftarrow \mathcal{K}; N_i \leftarrow \emptyset$
12 $(C = C') \wedge$	02 $\mathbb{S}, \mathbf{K}_c, \mathbf{K}_r \leftarrow \emptyset$
13 $\text{tup}((K_i, N, A, M), (K_j, N', A', M')) \wedge$	
14 $(\text{chk}(K_i, K_j) \vee \text{chk}(K_j, K_i))$	
<u>$\text{chk}(K_i, K_j)$</u>	procedure ENC(i, N, A, M)
16 if $\mathbf{t} = 00 \wedge K_i \notin \mathbf{K}_c \cup \mathbf{K}_r \wedge K_j \notin \mathbf{K}_c \cup \mathbf{K}_r$ then ret 1	20 if $K_i \notin \mathbf{K}_r \cup \mathbf{K}_c \wedge N \in \mathbf{N}_i$
17 if $\mathbf{t} = 01 \wedge K_i \notin \mathbf{K}_r \cup \mathbf{K}_c \wedge K_j \notin \mathbf{K}_c$ then ret 1	21 then ret \perp
18 if $\mathbf{t} = 0X \wedge K_i \notin \mathbf{K}_r \cup \mathbf{K}_c$ then ret 1	22 $C \leftarrow \Pi.E(K_i, N, A, M)$
19 if $\mathbf{t} = 11 \wedge K_i \notin \mathbf{K}_c \wedge K_j \notin \mathbf{K}_c$ then ret 1	23 $\mathbf{N}_i \leftarrow \mathbb{N}$
1A if $\mathbf{t} = 1X \wedge K_i \notin \mathbf{K}_c$ then ret 1	23 $\mathbb{S} \leftarrow \{(K_i, N, A, M, C)\}$
1B if $\mathbf{t} = XX$ then ret 1	24 ret C
1C ret 0	
<u>$\text{tup}((K_i, N, A, M), (K_j, N', A', M'))$</u>	procedure DEC(i, N, A, C)
1D $T_i, T_j \leftarrow (\perp, \perp, \perp, \perp)$	30 $M \leftarrow \Pi.D(K_i, N, A, C)$
1E if $\mathbf{u}[0] = 1$ then $T_i[0] \leftarrow K_i; T_j[0] \leftarrow K_j$	31 $\mathbb{S} \leftarrow \{(K_i, N, A, M, C)\}$
1F if $\mathbf{u}[1] = 1$ then $T_i[1] \leftarrow N; T_j[1] \leftarrow N'$	32 ret M
1G if $\mathbf{u}[2] = 1$ then $T_i[2] \leftarrow A; T_j[2] \leftarrow A'$	
1H if $\mathbf{u}[3] = 1$ then $T_i[3] \leftarrow M; T_j[3] \leftarrow M'$	
1I ret $T_i \neq T_j$	procedure REV(i)
	40 $\mathbf{K}_r \leftarrow \{K_i\}; \mathbf{ret} K_i$
	procedure COR(i, K)
	50 $K_i \leftarrow K; \mathbf{K}_c \leftarrow \{K_i\}$

Fig. 5: **An alternative CAE-security game.** This definition of CAE-security parametrizes the game with a four-bit string \mathbf{u} that dictates which elements—key, nonce, AD, message—should count as an adversarial win when misattributed.

B nAE-Security of CTX

PRIVACY AND AUTHENTICITY. We find it convenient to approach privacy and authenticity separately for CTX instead of using an all-in-one nAE definition. As such, we first recall games capturing these two notions [19]. Let Π be an nAE scheme. Let \mathcal{A} be a nonce-respecting adversary attacking the *privacy* of Π . Adversary \mathcal{A} is asked to distinguish between a “real” and “ideal” encryption oracle. The real oracle E is initialized with a key K sampled uniformly at random and uses it to output $\Pi.E(K, N, A, M)$ for any queries (N, A, M) that \mathcal{A} makes.

The ideal oracle $\$$ always outputs a uniformly random string of length $|M| + \tau$ where τ is the expansion of Π . The advantage of \mathcal{A} in this privacy game is then quantified as:

$$\mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{A}) = \Pr[K \leftarrow \mathcal{K}; \mathcal{A}^{E_K(\cdot, \cdot, \cdot)} \Rightarrow 1] - \Pr[\mathcal{A}^{\$(\cdot, \cdot, \cdot)} \Rightarrow 1].$$

The authenticity game asks that the nonce-respecting adversary \mathcal{A} *forges*. That is, \mathcal{A} is asked to output a tuple (N, A, C) such that $\Pi.\mathcal{D}(K, N, A, C) \neq \perp$ for some key K sampled uniformly at random upon initialization of the game. Adversary \mathcal{A} has access to an encryption oracle that performs encryption under K . Its authenticity advantage is defined as:

$$\mathbf{Adv}_{\Pi}^{\text{auth}}(\mathcal{A}) = \Pr[K \leftarrow \mathcal{K}; \mathcal{A}^{E_K(\cdot, \cdot, \cdot)} \text{ forges }].$$

To prevent trivial wins, if \mathcal{A} made an encryption query (N, A, M) that returned C , it may not output (N, A, C) as its forgery.

We now prove Theorem 3 from Section 4.

Proof. (1) We begin with the privacy part of the theorem. Let G_0 and G_1 be the games presented in Fig. 6. Note G_0 uses the boxed code whereas G_1 does not. Let G_2 be the game presented in Fig. 7. We claim that the advantage of adversary \mathcal{A}_1 is

$$\mathbf{Adv}_{\text{CTX}[\Pi]}^{\text{priv}}(\mathcal{A}_1) = \Pr[G_0(\mathcal{A}_1)] - \Pr[G_2(\mathcal{A}_2)] \quad (4)$$

where $\Pr[G(\mathcal{A})]$ denotes the probability that running adversary \mathcal{A} in game G results in the Finalize procedure of G returning true.

One can observe that G_0 is exactly the real privacy game of nAE security using CTX as the scheme. There are two tables keeping track of random oracle entries, HT and ET. The former records new random oracle (RO) entries generated by direct calls to H and the latter by calls to ENC. On the other hand, G_2 is the ideal privacy game for nAE security, coupled with access to a random oracle. Hence, Equation 4 follows by the definition of nAE privacy. Equation 4 is equivalent to the following:

$$\mathbf{Adv}_{\text{CTX}[\Pi]}^{\text{priv}}(\mathcal{A}_1) = \Pr[G_0(\mathcal{A}_1)] - \Pr[G_1(\mathcal{A}_1)] + \Pr[G_1(\mathcal{A}_1)] - \Pr[G_2(\mathcal{A}_1)] \quad (5)$$

From here, we can build adversary \mathcal{B}_1 such that

$$\Pr[G_1(\mathcal{A}_1)] - \Pr[G_2(\mathcal{A}_1)] \leq \mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{B}_1). \quad (6)$$

Adversary \mathcal{B}_1 behaves as follows. First, it initializes a table HT that it will maintain during its execution. Then, it runs adversary \mathcal{A}_1 . When \mathcal{A}_1 makes a query of:

- $\text{ENC}(N, A, M)$ - Adversary \mathcal{B}_1 calls its own encryption oracle with the same N, A, M , getting back $C \parallel T$ as a response. It then samples $T^* \leftarrow \{0, 1\}^\mu$ and returns $C \parallel T^*$ to \mathcal{A}_1 .

<p>Games $\boxed{G_0}/G_1$</p> <pre> procedure Initialize() 00 $K \leftarrow \mathcal{K}$ procedure ENC(N, A, M) 10 $C \parallel T \leftarrow \Pi \cdot \mathcal{E}_K(N, A, M)$ 11 $T^* \leftarrow \{0, 1\}^\mu$ 12 if $\text{HT}[K, N, A, T] \neq \perp$ then $\text{bad} \leftarrow \text{true}$; $T^* \leftarrow \text{HT}[K, N, A, T]$ 13 $\text{ET}[K, N, A, T] \leftarrow T^*$ 14 ret $C \parallel T^*$ procedure H(L, N, A, T) 20 if $\text{HT}[L, N, A, T] \neq \perp$ then ret $\text{HT}[L, N, A, T]$ 21 $\text{HT}[L, N, A, T] \leftarrow \{0, 1\}^\mu$ 22 if $\text{ET}[L, N, A, T] \neq \perp$ then 23 $\text{bad} \leftarrow \text{true}$; $\text{HT}[L, N, A, T] \leftarrow \text{ET}[L, N, A, T]$ 24 ret $\text{HT}[L, N, A, T]$ procedure Finalize(d) 30 ret d </pre>

Fig. 6: Games G_0 and G_1 for the privacy proof of Theorem 3. The two games are identical except G_0 contains the boxed code and G_1 does not.

- $H(L, N, A, T)$ - Adversary \mathcal{B}_1 checks to see if there is an entry in $\text{HT}[L, N, A, T]$. If not, then it samples a string uniformly at random from $\{0, 1\}^\mu$ and records it at $\text{HT}[L, N, A, T]$. It always responds to \mathcal{A}_1 with $\text{HT}[L, N, A, T]$.

When \mathcal{A}_1 terminates and outputs a bit, \mathcal{B}_1 outputs the same bit. When \mathcal{B}_1 's encryption oracle is fake, then it perfectly simulates G_2 as the ciphertext bodies C it returns to \mathcal{A}_1 will be uniformly random strings. When \mathcal{B}_1 's encryption oracle is real, then it perfectly simulates G_1 as C would be generated through encryption with Π under some hidden key (which is unknown to \mathcal{B}_1). This gives the advantage term of Equation 6.

Games G_0 and G_1 are identical-until-**bad**. By the Fundamental Lemma of Game Playing [6], we have that the advantage of an adversary distinguishing these two games is at most the probability of **bad** being set. That is, we have

$$\Pr[G_0(\mathcal{A}_1)] - \Pr[G_1(\mathcal{A}_1)] \leq \Pr[G_1(\mathcal{A}_1) \text{ sets } \text{bad}]. \quad (7)$$

In the games, **bad** gets set to true on lines 12 and 22 in Fig. 6 when the tables HT and ET are checked for an entry. Recall that the table ET records

```

Games  $G_2$ 

procedure ENC( $N, A, M$ )
40  $C \leftarrow \{0, 1\}^{|M|}$ ;  $T^* \leftarrow \{0, 1\}^\mu$ ; ret  $C \parallel T^*$ 

procedure H( $L, N, A, T$ )
50 if  $\text{HT}[L, N, A, T] = \perp$  then  $\text{HT}[L, N, A, T] \leftarrow \{0, 1\}^\mu$ 
51 ret  $\text{HT}[L, N, A, T]$ 

procedure Finalize( $d$ )
60 ret  $d$ 

```

Fig. 7: Games G_2 for the proof of Theorem 3.

mappings of K, N, A, T quadruples to random T^* that are generated during encryption. The key K here is fixed to the one sampled at game initialization. The table HT, on the other hand, tracks mappings from L, N, A, T quadruples to random T^* generated during random oracle queries. The key L here is part of the adversary's query. The flag **bad** gets set to true if one oracle (either encryption or the random oracle) finds an entry already recorded in the other's table when trying to generate the random T^* . For example, suppose the adversary makes a query (K, N, A, T) to H that results in some T^* . If the adversary later queries encryption with N, A, M such that $C \parallel T$ is the result of Π 's encryption, then the T^* returned to the adversary needs to be the T^* in the random oracle query. This is covered by the table HT. The other table ET covers the other direction—when a later RO query needs a tag from a previous encryption query.

Observe that for **bad** to be set true, the adversary will need to make a query to H with K , the secret key, as the argument L . It has to in order to satisfy the conditions of either 12 or 22. Following this, game G_3 in Fig. 8 is set up in a way such that an adversary wins if it queries the random oracle with the secret key. It runs its encryption oracle like G_1 . Hence, we have that

$$\Pr[G_1(\mathcal{A}_1) \text{ sets } \mathbf{bad}] \leq \Pr[G_3(\mathcal{A}_1)]. \quad (8)$$

Now, we build adversary \mathcal{B}_2 such that

$$\Pr[G_3(\mathcal{A}_1)] \leq \mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{B}_2) + \frac{q_H}{2^{\delta+\tau}} \quad (9)$$

Adversary \mathcal{B}_2 initializes with a table HT and a set S . It runs \mathcal{A}_1 and responds to encryption queries just like adversary \mathcal{B}_1 . For queries $H(L, N, A, T)$, \mathcal{B}_2 answers like \mathcal{B}_1 except it also adds L to its set S .

When \mathcal{A}_1 terminates, \mathcal{B}_2 ignores its output. It then picks a nonce N^* that was not used by \mathcal{A}_1 in any of its ENC queries. It picks a message $M^* \leftarrow \{0, 1\}^\delta$. Recall that $\delta \geq 0$ is the adjustable parameter from the theorem statement. Then

```

Games  $G_3$ 

procedure Initialize()
70  $K \leftarrow \mathcal{K}; S \leftarrow \emptyset$ 

procedure ENC( $N, A, M$ )
80  $C \parallel T \leftarrow \Pi.\mathcal{E}_K(N, A, M); T^* \leftarrow \{0, 1\}^\mu; \text{ret } C \parallel T^*$ 

procedure H( $L, N, A, T$ )
90 if  $\text{HT}[L, N, A, T] = \perp$  then  $\text{HT}[L, N, A, T] \leftarrow \{0, 1\}^\mu$ 
91  $S \leftarrow S \cup \{L\}; \text{ret } \text{HT}[L, N, A, T]$ 

procedure Finalize( $d$ )
AO ret ( $K \in S$ )

```

Fig. 8: Games G_3 for the proof of Theorem 3.

it queries its own ENC with N^*, ε, M^* and gets back a response C^* of the length $\delta + \tau$. Next, it sets a flag $b \leftarrow 0$ before executing a loop, iterating over every key $L \in S$. Each iteration, it checks whether $\Pi.\mathcal{E}_L(N^*, \varepsilon, M^*)$ outputs C^* , setting $b \leftarrow 1$ if one does. Finally, \mathcal{B}_2 outputs b as its output.

Let E_1 be the event that \mathcal{B}_2 outputs 1 in its real game and E_0 be the event it does so in its ideal game. Then, we have that $\Pr[E_1] \geq \Pr[G_3(\mathcal{A}_1)]$ because if the key K (the key to \mathcal{B}_2 's real game) is in the set S , then b is set to 1 when $L = K$ during \mathcal{B}_2 's loop. For E_0 , we have that $\Pr[E_0] \leq q_H/2^{\delta+\tau}$. Since S is fixed independently of the random C^* returned by the ideal encryption oracle, each iteration of \mathcal{B}_2 's loop sets b to 1 is at most $2^{-\delta+\tau}$. Applying the union bound across iterations, we get the bound for $\Pr[E_0]$. Finally, we get

$$\mathbf{Adv}_{\Pi}^{\text{priv}}(\mathcal{B}_2) = \Pr[E_1] - \Pr[E_0] \geq \Pr[G_3(\mathcal{A}_1)] - \frac{q_H}{2^{\delta+\tau}}$$

which gives Equation 9. Combining Equations 4, 6, and 9 yields the stated privacy bound in the theorem.

(2) We now proceed with the authenticity part of Theorem 3. Let \mathcal{A}_2 be the adversary attacking the authenticity of $\text{CTX}[\Pi]$. We can construct an adversary \mathcal{B}_3 attacking the authenticity of Π such that

$$\mathbf{Adv}_{\text{CTX}[\Pi]}^{\text{auth}}(\mathcal{A}_2) \leq \mathbf{Adv}_{\Pi}^{\text{auth}}(\mathcal{B}_3) + \frac{1}{2^\mu}. \quad (10)$$

We define \mathcal{B}_3 as follows. Adversary \mathcal{B}_3 responds to \mathcal{A}_2 's ENC and H queries the same way that \mathcal{B}_1 does. When \mathcal{A}_2 terminates with a forgery $(N_2, A_2, C_2 \parallel T_2)$, adversary \mathcal{B}_3 then picks a message M^* of length δ , and a nonce N^* that has not been used by \mathcal{A}_2 in its ENC queries and is not the nonce in \mathcal{A}_2 's forgery

$(N^* \neq N_2)$. It then makes a query to its own encryption oracle ENC with N^*, ε, M^* , getting back a response C^* .

Now \mathcal{B}_3 selects particular entries in its table HT. We write these entries as quintuples (L, N, A, T, T^*) to succinctly denote the mapping of (L, N, A, T) to T^* . Specifically, \mathcal{B}_3 picks entries (L, N, A, T, T^*) such that $N = N_2$, $A = A_2$, and $T^* = T_2$. It then iterates through all such entries and tests whether $\Pi.\mathcal{E}_L(N, A, M^*) = C^*$. Let S be the set of entries that satisfy this condition. Adversary \mathcal{B}_3 then executes the following:

```

for  $(L, N, A, T, T^*) \in S$  do
     $M \leftarrow \Pi.\mathcal{D}_1(L, N, A, C_2)$ 
     $T' \leftarrow \Pi.\mathcal{E}_2(L, N, A, M)$ 
    if  $T = T'$  then ret  $(N, A, C_2 \parallel T)$ 

```

Recall that \mathcal{E}_2 and \mathcal{D}_1 are the algorithms of a tag-based nAE scheme that produce the authentication tag and decrypt the ciphertext core respectively. This loop checks which of the candidate entries contains a forgery \mathcal{B}_3 can use for C_2 by verifying a tag (for the scheme Π) for it. If the loop finishes without successfully verifying a tag, this means that \mathcal{A}_2 failed its own forgery, which would cause \mathcal{B}_3 to fail as well. When \mathcal{A}_2 succeeds in forging, then \mathcal{B}_3 can recover a tag T appropriate for its own forgery. This gives us the authenticity advantage term for \mathcal{B}_3 in the bound.

It is possible for \mathcal{A}_2 to forge without \mathcal{B}_3 being able to recover the tag it needs. Suppose \mathcal{A}_2 's forgery is $(N, A, C \parallel T^*)$. Then, for example, \mathcal{A}_2 could never ask H the appropriate query to fill in HT with T^* . However, all responses to H queries are independent of one another, so one response for H tells \mathcal{A}_2 nothing about another response for H. This is also true for the random tags generated by any ENC queries made by \mathcal{A}_2 . So, without querying to H the exact (L, N, A, T) used by \mathcal{B}_3 's encryption oracle to make C , the best \mathcal{A}_2 can do to come up with a valid T^* for its forgery is by guessing. For a fixed key, nonce, AD, and ciphertext core, there is exactly one tag that verifies. Following this, the probability \mathcal{A}_2 creates a valid tag T^* and forges by guessing is $1/2^\mu$ as it has a single guess for its forgery over 2^μ possible tags. This gives us the last term in the bound. \square

C CAE_{ox} Attack on CTX

Let Π' be a tag-based nAE scheme that uses a PRG G to generate a one time pad for the plaintext M using the key K and the nonce N to get a ciphertext core C and a tagging function $\mathcal{E}_2(K, N, A, M)$ for an AD A to produce a tag T . It returns $C \parallel T$ as its final ciphertext. AES-GCM is an example of such a scheme. Now let Π be a scheme that runs exactly the same as Π' except if the key input is a special reserved key, say $K^* = 0^k$ where k is the key length of Π , then \mathcal{E}_2 outputs a tag $T = 0^t$ for all N, A, M where t is the tag length of Π .

We give an adversary \mathcal{A} against $\text{CTX}[\Pi, H]$ even if H is a collision-resistant hash function. Let μ be the output length of H and q be the number of encryption queries \mathcal{A} makes. Adversary \mathcal{A} fixes N_1, \dots, N_q distinct nonces as well as some

message M and some AD A . It then makes q queries to its encryption oracle where the i th query is in the form $C_i \parallel T_i \leftarrow \text{ENC}(N_i, A, M)$. Note that these queries are made under some hidden key K where K is the target honest key that \mathcal{A} is trying to collide with.

Now \mathcal{A} chooses a fresh nonce N distinct from its previous nonces and chooses distinct AD values A_1, \dots, A_q that are different from the previous A . It then computes a q candidate tags where the j th hash computation is in the form $U_j \leftarrow H(0^k, N, A_j, 0^t)$.

By the birthday paradox, with probability $\Omega(q^2/2^\mu)$, there are indices i, j such that $T_i = U_j$. Now \mathcal{A} computes $M_j \leftarrow G(0^k, N) \oplus C_i$. We end up with the misattribution $\Pi.\mathcal{E}(0^k, N, A_j, M_j) = C_i \parallel T_i$ through the following equations:

$$\begin{aligned} \Pi.\mathcal{E}(0^k, N, A_j, M_j) &= G(0^k, N) \oplus M_j \parallel H(0^k, N, A_j, \mathcal{E}_2(0^k, N, A_j, M_j)) \\ &= G(0^k, N) \oplus G(0^k, N) \oplus C_i \parallel H(0^k, N, A_j, 0^t) \quad // \text{ def of } M_j \& \mathcal{E}_2 \\ &= C_i \parallel U_j = C_i \parallel T_i \quad // \text{ by def of } U_j \text{ and birthday collision} \end{aligned}$$

So, CTX is unable to avoid a birthday bound even for CAE_{ox}-security. It is possible that it may be able to do so with additional assumptions on the nAE scheme it uses. We leave the investigation of such assumptions and the possibility of other cAE schemes with better CAE_{ox}-security to future work.